



(11) Publication number : **0 614 977 A2**

(12)

EUROPEAN PATENT APPLICATION

(21) Application number : **94301587.5**

(22) Date of filing : **07.03.94**

(51) Int. Cl.⁵ : **C12N 15/12, C07K 13/00, C12N 1/21, C12N 5/10, C07K 15/28, C12N 5/16, C12Q 1/68, A61K 37/02, A61K 48/00, C12P 21/08**

(30) Priority : **05.03.93 US 27498**
01.07.93 US 85000

(43) Date of publication of application :
14.09.94 Bulletin 94/37

(84) Designated Contracting States :
AT BE CH DE DK ES FR GB GR IE IT LI LU MC NL PT SE

(71) Applicant : **THE GENERAL HOSPITAL CORPORATION**
55 Fruit Street
Boston, MA 02114 (US)

(72) Inventor : **MacDonald, Marcy E.**
462 Waltham Street
Lexington, Massachusetts 02173 (US)
Inventor : **Ambrose, Christine M.**
42-8th Street, No. 3105 Charlestown
Massachusetts 02129 (US)
Inventor : **Duyao, Mabel P.**
24 Aberdeen Avenue
Cambridge, Massachusetts 02138 (US)
Inventor : **Gusella, James F.**
7 Woodstock Drive
Framingham, Massachusetts 01701 (US)

(74) Representative : **Wright, Simon Mark et al**
Kilburn & Strode
30 John Street
London WC1N 2DD (GB)

(54) **Huntingtin DNA, protein and uses thereof.**

(57) A novel gene, *huntingtin*, is described, encoding huntingtin protein, recombinant vectors and hosts capable of expressing huntingtin. Methods for the diagnosis and treatment of Huntington's disease are also provided.

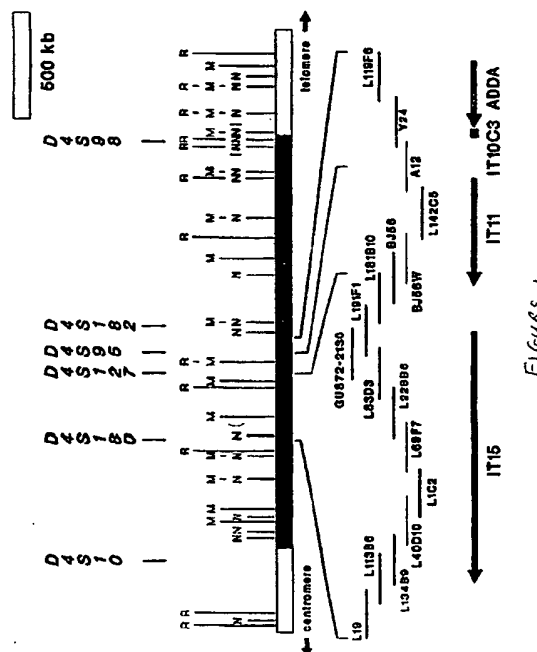


FIGURE 1

Field of the Invention

The invention is in the field of the detection and treatment of genetic diseases. Specifically, the invention is directed to the *huntingtin* gene (also called the IT15 gene), huntingtin protein encoded by such gene, and the use of this gene and protein in assays (1) for the detection of a predisposition to develop Huntington's disease, (2) for the diagnosis of Huntington's disease (3) for the treatment of Huntington's disease, and (4) for monitoring the course of treatment of such treatment.

Background of the Invention

Huntington's disease (HD) is a progressive neurodegenerative disorder characterized by motor disturbance, cognitive loss and psychiatric manifestations (Martin and Gusella, *N. Engl. J. Med.* 315:1267-1276 (1986)). It is inherited in an autosomal dominant fashion, and affects about 1/10,000 individuals in most populations of European origin (Harper, P.S. *et al.*, in *Huntington's disease*, W.B. Saunders, Philadelphia, 1991). The hallmark of HD is a distinctive choreic movement disorder that typically has a subtle, insidious onset in the fourth to fifth decade of life and gradually worsens over a course of 10 to 20 years until death. Occasionally, HD is expressed in juveniles typically manifesting with more severe symptoms including rigidity and a more rapid course. Juvenile onset of HD is associated with a preponderance of paternal transmission of the disease allele. The neuropathology of HD also displays a distinctive pattern, with selective loss of neurons that is most severe in the caudate and putamen regions of the brain. The biochemical basis for neuronal death in HD has not yet been explained, and there is consequently no treatment effective in delaying or preventing the onset and progression of this devastating disorder.

The genetic defect causing HD was assigned to chromosome 4 in 1983 in one of the first successes of linkage analysis using polymorphic DNA markers in man (Gusella *et al.*, *Nature* 306:234-238 (1983)). Since that time, we have pursued a location cloning approach to isolating and characterizing the HD gene based on progressively refining its localization (Gusella, *FASEB J.* 3:2036-2041 (1989); Gusella, *Adv. Hum. Genet.* 20:125-151 (1991)). Among other work, this has involved the generation of new genetic markers in the region by a number of techniques (Pohl *et al.*, *Nucleic Acids Res.* 16:9185-9198 (1988); Whaley *et al.*, *Somat. Cell. Mol. Genet.* 17:83-91 (1991); MacDonald *et al.*, *J. Clin. Inv.* 84:1013-1016 (1989)), the establishment of genetic (MacDonald *et al.*, *Neuron* 3:183-190 (1989); Allitto *et al.*, *Genomics* 9:104-112 (1991)) and physical maps of the implicated regions (Bucan *et al.*, *Genomics* 6:1-15 (1990); Bates *et al.*, *Nature Genet.* 1:180-187 (1992); Doucette-Stamm *et al.*, *Somat. Cell Mol. Genet.* 17:471-480 (1991); Altherr *et al.*, *Genomics* 13:1040-1046 (1992)), the cloning of the 4p telomere of an HD chromosome in a YAC clone (Bates *et al.*, *Am. J. Hum. Genet.* 46:762-775 (1990); Youngman *et al.*, *Genomics* 14:350-356 (1992)), the establishment of YAC [yeast artificial chromosome] (Bates *et al.*, *Nature Genet.* 1:180-187 (1992)) and cosmid (Baxendale *et al.*, in preparation) contigs (a series of overlapping clones which together form a whole sequence) of the candidate region, as well as the analysis and characterization of a number of candidate genes from the region (Thompson *et al.*, *Genomics* 11:1133-1142 (1991); Taylor *et al.*, *Nature Genet.* 2:223-227 (1992); Ambrose *et al.*, *Hum. Mol. Genet.* 1:697-703 (1992)). Analysis of recombination events in HD kindreds has identified a candidate region of 2.2 Mb, between D4S10 and D4S98 in 4p16.3, as the most likely position of the HD gene (MacDonald *et al.*, *Neuron* 3:183-190 (1989); Bates *et al.*, *Am. J. Hum. Genet.* 49:7-16 (1991); Snell *et al.*, *Am. J. Hum. Genet.* 51:357-362 (1992)). Investigations of linkage disequilibrium between HD and DNA markers in 4p16.3 (Snell *et al.*, *J. Med. Genet.* 26:673-675 (1989); Theilman *et al.*, *J. Med. Genet.* 26:676-681 (1989)) have suggested that multiple mutations have occurred to cause the disorder (MacDonald *et al.*, *Am. J. Hum. Genet.* 49:723-734 (1991)). However, haplotype analysis using multi-allele markers has indicated that at least 1/3 of HD chromosomes are ancestrally related (MacDonald *et al.*, *Nature Genet.* 1: 99-103 (1992)). The haplotype shared by these HD chromosomes points to a 500 kb segment between D4S180 and D4S182 as the most likely site of the genetic defect.

Targeting this 500 kb region for saturation with gene transcripts, exon amplification has been used as a rapid method for obtaining candidate coding sequences (Buckler *et al.*, *Proc. Natl. Acad. Sci. USA* 88:4005-4009 (1991)). This strategy has previously identified three genes: the α -adducin gene (ADDA) (Taylor *et al.*, *Nature Genet.* 2:223-227 (1992)); a putative novel transporter gene (IT10C3) in the distal portion of this segment; and a novel G protein-coupled receptor kinase gene (IT11) in the central portion (Ambrose *et al.*, *Hum. Mol. Genet.* 1:697-703 (1992)). However, no defects implicating any of these genes as the HD locus have been found.

Summary of the Invention

A large gene, termed herein "huntingtin" or "IT15," has been identified that spans about 210 kb and encodes a previously undescribed protein of about 348 kDa. The huntingtin reading frame contains a polymorphic (CAG)_n trinucleotide repeat with at least 17 alleles in the normal population, varying from 11 to about 34 CAG copies. On HD chromosomes, the length of the trinucleotide repeat is substantially increased, for example, about 37 to at least 73 copies, and shows an apparent correlation with age of onset, the longest segments are detected in juvenile HD cases. The instability in length of the repeat is reminiscent of similar trinucleotide repeats in the fragile X syndrome and in myotonic dystrophy (Suthers *et al.*, *J. Med. Genet.* 29:761-765 (1992)). The presence of an unstable, expandable trinucleotide repeat on HD chromosomes in the region of strongest linkage disequilibrium with the disorder suggests that this alteration underlies the dominant phenotype of HD, and that huntingtin encodes the HD gene.

The invention is directed to the protein huntingtin, DNA and RNA encoding this protein, and uses thereof.

Accordingly, in a first embodiment, the invention is directed to purified preparations of the protein huntingtin, preferably substantially cell-free.

In a further embodiment, the invention is directed to a recombinant construct containing DNA or RNA encoding huntingtin.

In a further embodiment, the invention is directed to a vector containing such huntingtin-encoding nucleic acid.

In a further embodiment, the invention is directed to a host transformed with such vector.

In a further embodiment, the invention is directed to a method for producing huntingtin from such recombinant host.

In a further embodiment, the invention is directed to a method for diagnosing Huntington's disease using such huntingtin DNA, RNA and/or protein.

In a further embodiment, the invention is directed to a method for treating Huntington's disease using such huntingtin DNA, RNA and/or protein.

In a further embodiment, the invention is directed to a method of gene therapy of a symptomatic or pre-symptomatic patient, such method comprising providing a functional *huntingtin* gene with a (CAG)_n repeat of the normal range of 11-34 copies to the desired cells of such patient in need of such treatment, in a manner that permits the expression of the huntingtin protein provided by such gene, for a time and in a quantity sufficient to provide the huntingtin function to the cells of such patient.

In a further embodiment, the invention is directed to a method of gene therapy of a symptomatic or pre-symptomatic patient, such method comprising providing a functional *huntingtin* antisense gene to the desired cells of such patient in need of such treatment, in a manner that permits the expression of huntingtin antisense RNA provided by such gene, for a time and in a quantity sufficient to inhibit huntingtin mRNA expression in the cells of such patient.

In a further embodiment, the invention is directed to a method of gene therapy of a symptomatic or pre-symptomatic patient, such method comprising providing a functional *huntingtin* gene to the cells of such patient in need of such gene; in one embodiment the functional huntingtin gene contains a (CAG)_n repeat size between 11-34 copies.

In a further embodiment, the invention is directed to a method for diagnosing Huntington's disease or a predisposition to develop Huntington's disease in a patient, such method comprising determining the number of (CAG)_n repeats present in the huntingtin gene in such patient and especially in the affected tissue of such patient.

In a further embodiment, the invention is directed to a method for treating Huntington's disease in a patient, such method comprising decreasing the number of huntingtin (CAG)_n repeats in the huntingtin gene in the desired cells of such patient.

Brief Description of the Drawings

FIGURE 1. Long-range restriction map of the HD candidate region. A partial long range restriction map of 4p16.3 is shown (adapted from Lin *et al.*, *Somat. Cell Mol. Genet.* 17:481-488 (1991)). The HD candidate region determined by recombination events is depicted as a hatched line between D4S10 and D4S98. The portion of the HD candidate region implicated as the site of the defect by linkage disequilibrium haplotype analysis (MacDonald *et al.*, *Nature Genet.* 1:99-103 (1992)) is shown as a filled box. Below the map schematic, the region from D4S180 to D4S182 is expanded to show the cosmid contig (averaging 40 kb/cosmid). The genomic coverage and where known the transcriptional orientation (arrow 5' to 3') of the huntingtin (IT15), IT11, IT10C3 and ADDA genes is also shown. Locus names above the map denote selected polymorphic markers that have

been used in HD families. The positions of *D4S127* and *D4S95* which form the core of haplotype in the region of maximum disequilibrium are also shown in the cosmid contig. Restriction sites are given for Not I (N), Mlu I (M) and Nru I (R). Sites displaying complete digestion are shown in boldface while sites subject to frequent incomplete digestion are shown as lighter symbols. Brackets around the "N" symbols indicate the presence of additional clustered Not I sites.

FIGURE 2. Northern blot analysis of the huntingtin (IT15) transcript. Results of the hybridization of IT15A to a Northern blot of RNA from normal (lane 1) and HD homozygous (lane 2 and 3) lymphoblasts are shown. A single RNA of about 11 kb was detected in all three samples, with slight apparent variations being due to unequal RNA concentrations. The HD homozygotes are independent, deriving from the large an American family (lane 2) and the large Venezuelan family (lane 3), respectively. The Venezuelan HD chromosome has a 4p16.3 haplotype of "5 2 2" defined by a (GT)_n polymorphism at *D4S127* and VNTR and TaqI RFLPs at *D4S95*. The American homozygote carries the most common 4p16.3 haplotype found on HD chromosomes: "2 11 1" (MacDonald *et al.*, *Nature Genet.* 1:99-103 (1992)).

FIGURE 3. Schematic of cDNA clones defining the IT15 transcript. Five cDNAs are represented under a schematic of the composite IT15 sequence. The thin line corresponds to untranslated regions. The thick line corresponds to coding sequence, assuming initiation of translation at the first Met codon in the open reading frame. Stars mark the positions of the following exon clones 5' to 3': DL83D3-8, DL83D3-1, DL228B6-3, DL228B6-5, DL228B6-13, DL69F7-3, DL178H4-6, DL118F5-U and DL134B9-U4.

The composite sequence was derived as follows. From 22 bases 3' to the putative initiator Met ATG, the sequence was compiled from the cDNA clones and exons shown. There are 9 bases of sequence intervening between the 3' end of IT16B and the 5' end of IT15B. These were by PCR amplification of first strand cDNA and sequencing of the PCR product. At the 5' end of the composite sequence, the cDNA clone IT16C terminates 27 bases upstream of the (CAG)_n. However, when IT16C was identified, we had already generated genomic sequence surrounding the (CAG)_n in an attempt to generate new polymorphisms. This sequence matched the IT16C sequence, and extended it 337 bases upstream, including the apparent Met initiation codon.

FIGURE 4. Composite sequence of huntingtin (IT15)(SEQ ID NO:5 and SEQ ID NO:6). The composite DNA sequence of huntingtin (IT15) is shown (SEQ ID NO:5). The predicted protein product (SEQ ID NO:6) is shown below the DNA sequence, based on the assumption that translation begins at the first in-frame methionine of the long open reading frame.

FIGURE 5. DNA sequence analysis of the (CAG)_n repeat. DNA sequence shown in panels 1, 2 and 3, demonstrates the variation in the (CAG)_n repeat detected in normal cosmid L191F1 (1), cDNA IT16C (2), and HD cosmid GUS72-2130. Panels 1 and 3 were generated by direct sequencing of cosmid subclones using the following primer (SEQ ID NO:1):

5' GGC GGG AGA CCG CCA TGG CG 3'.

Panel 2 was generated using the pBSKII T7 primer (SEQ ID NO:2):

5' AAT ACG ACT CAC TAT AG 3'.

FIGURE 6. PCR analysis of the (CAG)_n repeat in a Venezuelan HD sibship with some offspring displaying juvenile onset. Results of PCR analysis of a sibship in the Venezuela HD pedigree are shown. Affected individuals are represented by shaded symbols. Progeny are shown as triangles for confidentiality. AN1, AN2 and AN3 mark the positions of the allelic products from normal chromosomes. AE marks the range of PCR products from the HD chromosome. The intensity of background constant bands, which represent a useful reference for comparison of the above PCR products, varies with slight differences in PCR conditions. The PCR products from cosmids L191F1 and GUS72-2130 are loaded in lanes 12 and 13 and have 18 and 48 CAG repeats, respectively.

FIGURE 7. PCR analysis of the (CAG)_n repeat in a Venezuelan HD sibship with offspring homozygous for the same HD haplotype. Results of PCR analysis of a sibship from the Venezuela HD pedigree in which both parents are affected by HD are shown. Progeny are shown as triangles for confidentiality and no HD diagnostic information is given to preserve the blind status of investigators in the Venezuelan Collaborative Group. AN1 and AN2 mark the positions of the allelic products from normal parental chromosomes. AE marks the range of PCR products from the HD chromosome. The PCR products from cosmids L191F1 and GUS72-2130 are loaded in lanes 29 and 30 and have 18 and 48 CAG repeats, respectively.

FIGURE 8. PCR analysis of the (CAG)_n repeat in members of an American family with an individual homozygous for the major HD haplotype. Results of PCR analysis of members of an American family segregating the major HD haplotype. AN marks the range of normal alleles; AE marks the range of HD alleles. Lanes 1, 3,

4, 5, 7 and 8 represent PCR products from related *HD* heterozygotes. Lane 2 contains the PCR products from a member of the family homozygous for the same *HD* chromosome. Lane 6 contains PCR products from a normal individual. Pedigree relationships and affected status are not presented to preserve confidentiality. The PCR products from cosmids L191F1 and GUS72-2130 (which was derived from the individual represented in lane 2) are loaded in lanes 9 and 10 and have 18 and 48 CAG repeats, respectively.

FIGURES 9 and 10. PCR analysis of the (CAG)_n repeat in two families with supposed new mutation causing *HD*. Results of PCR analysis of two families in which sporadic *HD* cases representing putative new mutants are shown. Individuals in each pedigree are numbered by generation (Roman numerals) and order in the pedigree. Triangles are used to protect confidentiality. Filled symbols indicate symptomatic individuals. The different chromosomes segregating in the pedigree have been distinguished by extensive typing with polymorphic markers in 4p16.3 and have been assigned arbitrary numbers shown above the gel lanes. The starred chromosomes (3 in Figure 9, 1 in Figure 10) represent the presumed *HD* chromosome. AN denotes the range of normal alleles; AE denotes the range of alleles present in affected individuals and in their unaffected relatives bearing the same chromosomes.

FIGURE 11. Comparison of (CAG)_n Repeat Unit Number on Control and *HD* Chromosomes. Frequency distributions are shown for the number of (CAG)_n repeat units observed on 425 *HD* chromosomes from 150 independent families, and from 545 control chromosomes.

FIGURE 12. Comparison of (CAG)_n Repeat Unit Number on Maternally and Paternally Transmitted *HD* Chromosomes. Frequency distributions are shown for the 134 and 161 *HD* chromosomes from Figure 11 known to have been transmitted from the mother (Panel A) and father (Panel B), respectively. The two distributions differ significantly based on a t-test ($t_{272.3} = 5.34$, $p < 0.0001$).

FIGURE 13. Comparison of (CAG)_n Repeat Unit Number on *HD* Chromosomes from Three Large Families with Different *HD* Founders. Frequency distributions are shown for 75, 25 and 35 *HD* chromosomes from the Venezuelan *HD* family (Panel A) (Gusella, J.F., *et al.*, *Nature* 306:234-238 (1983); Wexler, N.S., *et al.*, *Nature* 326:194-197 (1987)), Family Z (Panel B) and Family D (Panel C) (Folstein, S.E., *et al.*, *Science* 229:776-779 (1985)), respectively. The Venezuelan distribution did not differ from the overall *HD* chromosome distribution in Figure 11 ($t_{79.7} = 1.58$, $p < 0.12$). Both Family Z and Family D did produce distributions significantly different from the overall *HD* distribution ($t_{42.2} = 6.73$, $p < 0.0001$ and $t_{458} = 2.90$, $p < 0.004$, respectively).

Figure 14. Relationship of (CAG)_n Repeat Length in Parents and Corresponding Progeny. Repeat length on the *HD* chromosome in mothers (Panel A) or fathers (Panel B) is plotted against the repeat length in the corresponding offspring. A total of 25 maternal transmissions and 37 paternal transmissions were available for typing.

FIGURE 15. Amplification of the *HD* (CAG)_n Repeat From Sperm and Lymphoblast DNA. DNA from sperm (S) and lymphoblasts (L) for 5 members (pairs 1-5) of the Venezuelan *HD* pedigree aged 24-30 were used for PCR amplification of the *HD* (CAG)_n repeat. The lower band in each lane derives from the normal chromosome.

FIGURE 16. Relationship of Repeat Unit Length with Age of Onset. Age of onset was established for 234 diagnosed *HD* gene carriers and plotted against the repeat length observed on both the *HD* and normal chromosomes in the corresponding lymphoblast lines.

Detailed Description of the Invention

In the following description, reference will be made to various methodologies known to those of skill in the art of molecular genetics and biology. Publications and other materials setting forth such known methodologies to which reference is made are incorporated herein by reference in their entireties as though set forth in full.

The IT15 gene described herein is a gene from the proximal portion of the 500 kb segment between human chromosome 4 markers *D4S180* and *D4S182*. The huntingtin gene spans about 210 kb of DNA and encodes a previously undescribed protein of about 348 kDa. The huntingtin reading frame contains a polymorphic (CAG)_n trinucleotide repeat with at least 17 alleles in the normal human population, where the repeat number varies from 11 to about 34 CAG copies in such alleles. This is the gene of the human chromosome that, as shown herein, suffers the presence of an unstable, expanded number of CAG trinucleotide repeats in Huntington's disease patients, such that the number of CAG repeats in the huntingtin gene increases to a range of 37 to at least 86 copies. These results are the basis of a conclusion that the huntingtin gene encodes a protein called "huntingtin," and that in such huntingtin gene the increase in the number of CAG repeats to a range of greater than about 37 repeats is the alteration that underlies the dominant phenotype of Huntington's disease. As used herein huntingtin gene is also called the Huntington's disease gene.

It is to be understood that the description below is applicable to any gene in which a CAG repeat within the gene is amplified in an aberrant manner resulting in a change in the regulation, localization, stability or translatability of the mRNA containing such amplified CAG repeat that is transcribed from such gene.

1. Cloning Of Huntingtin DNA And Expression Of Huntingtin Protein

The identification of huntingtin DNA and protein as the altered gene in Huntington's disease patients is exemplified below. In addition to utilizing the exemplified methods and results for the identification of deletions of the *huntingtin* gene in Huntington's disease patients, and for the isolation of the native human *huntingtin* gene, the sequence information presented in Figure 4 represents a nucleic acid and protein sequence, that, when inserted into a linear or circular recombinant nucleic acid construct such as a vector, and used to transform a host cell, will provide copies of *huntingtin* DNA and huntingtin protein that are useful sources for the native *huntingtin* DNA and huntingtin protein for the methods of the invention. Such methods are known in the art and are briefly outlined below.

The process for genetically engineering the *huntingtin* coding sequence, for expression under a desired promoter, is facilitated through the cloning of genetic sequences which are capable of encoding such huntingtin protein. Such cloning technologies can utilize techniques known in the art for construction of a DNA sequence encoding the huntingtin protein, such as, for example, polymerase chain reaction technologies utilizing the *huntingtin* sequence disclosed herein to isolate the *huntingtin* gene anew, or an allele thereof that varies in the number of CAG repeats in such gene, or polynucleotide synthesis methods for constructing the nucleotide sequence using chemical methods. Expression of the cloned *huntingtin* DNA provides huntingtin protein.

As used herein, the term "genetic sequences" is intended to refer to a nucleic acid molecule of DNA or RNA, preferably DNA. Genetic sequences that are capable of being operably linked to DNA encoding huntingtin protein, so as to provide for its expression and maintenance in a host cell are obtained from a variety of sources, including commercial sources, genomic DNA, cDNA, synthetic DNA, and combinations thereof. Since the genetic code is universal, it is to be expected that any DNA encoding the huntingtin amino acid sequence of the invention will be useful to express huntingtin protein in any host, including prokaryotic (bacterial) hosts, eukaryotic hosts (plants, mammals (especially human), insects, yeast, and especially any cultured cell populations).

If it is desired to select anew a gene encoding huntingtin from a library that is thought to contain a *huntingtin* gene, such library can be screened and the desired gene sequence identified by any means which specifically selects for a sequence coding for the *huntingtin* gene or expressed huntingtin protein such as, for example, a) by hybridization (under stringent conditions for DNA:DNA hybridization) with an appropriate *huntingtin* DNA probe(s) containing a sequence specific for the DNA of this protein, such sequence being that provided in Figure 4 or a functional derivative thereof that is, a shortened form that is of sufficient length to identify a clone containing the *huntingtin* gene, or b) by hybridization-selected translational analysis in which native *huntingtin* mRNA which hybridizes to the clone in question is translated *in vitro* and the translation products are further characterized for the presence of a biological activity of huntingtin, or c) by immunoprecipitation of a translated huntingtin protein product from the host expressing the huntingtin protein.

When a human allele does not encode the identical sequence to that of Figure 4, it can be isolated and identified as being *huntingtin* DNA using the same techniques used herein, and especially PCR techniques to amplify the appropriate gene with primers based on the sequences disclosed herein. Many polymorphic probes useful in the fine localization of genes on chromosome 4 are known and available (see, for example, "ATCC/NIH Repository Catalogue of Human and Mouse DNA Probes and Libraries," fifth edition, 1991, pages 4-6. For example, a useful *D4S10* probe is done designation pTV20 (ATCC 57605 and 57604); H5.52 (ATCC 61107 and 61106) and F5.53 (ATCC 61108).

Human chromosome 4-specific libraries are known in the art and available from the ATCC for the isolation of probes ("ATCC/NIH Repository Catalogue of Human and Mouse DNA Probes and Libraries," fifth edition, 1991, pages 72-73), for example, LL04NS01 and LL04NS02 (ATCC 57719 and ATCC57718) are useful for these purposes.

It is not necessary to utilize the exact vector constructs exemplified in the invention; equivalent vectors can be constructed using techniques known in the art. For example, the sequence of the huntingtin DNA is provided herein, (see Figure 4) and this sequence provides the specificity for the *huntingtin* gene; it is only necessary that a desired probe contain this sequence, or a portion thereof sufficient to provide a positive indication of the presence of the *huntingtin* gene.

Huntingtin genomic DNA may or may not include naturally occurring introns. Moreover, such genomic DNA can be obtained in association with the native *huntingtin* 5' promoter region of the gene sequences and/or with the native *huntingtin* 3' transcriptional termination region.

Such *huntingtin* genomic DNA can also be obtained in association with the genetic sequences which encode the 5' non-translated region of the *huntingtin* mRNA and/or with the genetic sequences which encode the *huntingtin* 3' non-translated region. To the extent that a host cell can recognize the transcriptional and/or translational regulatory signals associated with the expression of *huntingtin* mRNA and protein, then the

5' and/or 3' non-transcribed regions of the native *huntingtin* gene, and/or, the 5' and/or 3' non-translated regions of the huntingtin mRNA can be retained and employed for transcriptional and translational regulation.

Genomic DNA can be extracted and purified from any host cell, especially a human host cell possessing chromosome 4, by means well known in the art. Genomic DNA can be shortened by means known in the art, such as physical shearing or restriction digestion, to isolate the desired *huntingtin* gene from a chromosomal region that otherwise would contain more information than necessary for the utilization of the *huntingtin* gene in the hosts of the invention. For example, restriction digestion can be utilized to cleave the full-length sequence at a desired location. Alternatively, or in addition, nucleases that cleave from the 3'-end of a DNA molecule can be used to digest a certain sequence to a shortened form, the desired length then being identified and purified by polymerase chain reaction technologies, gel electrophoresis, and DNA sequencing. Such nucleases include, for example, Exonuclease III and *Bal31*. Other nucleases are well known in the art.

Alternatively, if it is known that a certain host cell population expresses huntingtin protein, then cDNA techniques known in the art can be utilized to synthesize a cDNA copy of the huntingtin mRNA present in such population.

For cloning the genomic or cDNA nucleic acid that encodes the amino acid sequence of the huntingtin protein into a vector, the DNA preparation can be ligated into an appropriate vector. The DNA sequence encoding huntingtin protein can be inserted into a DNA vector in accordance with conventional techniques, including blunt-ending or staggered-ending termini for ligation, restriction enzyme digestion to provide appropriate termini, filling in of cohesive ends as appropriate, alkaline phosphatase treatment to avoid undesirable joining, and ligation with appropriate ligases. Techniques for such manipulations are well known in the art.

When the huntingtin DNA coding sequence and an operably linked promoter are introduced into a recipient eukaryotic cell (preferably a human host cell) as a non-replicating, non-integrating, molecule, the expression of the encoded huntingtin protein can occur through the transient (nonstable) expression of the introduced sequence.

Preferably the coding sequence is introduced on a DNA molecule, such as a closed circular or linear molecule that is capable of autonomous replication. If integration into the host chromosome is desired, it is preferable to use a linear molecule. If stable maintenance of the *huntingtin* gene is desired on an extrachromosomal element, then it is preferable to use a circular plasmid form, with the appropriate plasmid element for autonomous replication in the desired host.

The desired gene construct, providing a gene coding for the huntingtin protein, and the necessary regulatory elements operably linked thereto, can be introduced into a desired host cells by transformation, transfection, or any method capable of providing the construct to the host cell. A marker gene for the detection of a host cell that has accepted the *huntingtin* DNA can be on the same vector as the *huntingtin* DNA or on a separate construct for cotransformation with the huntingtin coding sequence construct into the host cell. The nature of the vector will depend on the host organism.

Suitable selection markers will depend upon the host cell. For example, the marker can provide biocide resistance, e.g., resistance to antibiotics, or heavy metals, such as copper, or the like.

Factors of importance in selecting a particular plasmid or viral vector include: the ease with which recipient cells that contain the vector can be recognized and selected from those recipient cells which do not contain the vector; the number of copies of the vector which are desired in a particular host; and whether it is desirable to be able to "shuttle" the vector between host cells of different species.

When it is desired to use *S. cerevisiae* as a host for a shuttle vector, preferred *S. cerevisiae* yeast plasmids include those containing the 2-micron circle, etc., or their derivatives. Such plasmids are well known in the art and are commercially available.

Oligonucleotide probes specific for the *huntingtin* sequence can be used to identify clones to huntingtin and can be designed *de novo* from the knowledge of the amino acid sequence of the protein as provided herein in Figure 4 or from the knowledge of the nucleic acid sequence of the DNA encoding such protein as provided herein in Figure 4 or of a related protein. Alternatively, antibodies can be raised against the huntingtin protein and used to identify the presence of unique protein determinants in transformants that express the desired cloned protein.

A nucleic acid molecule, such as DNA, is said to be "capable of expressing" a huntingtin protein if that nucleic acid contains expression control sequences which contain transcriptional regulatory information and such sequences are "operably linked" to the huntingtin nucleotide sequence which encode the huntingtin polypeptide.

An operable linkage is a linkage in which a sequence is connected to a regulatory sequence (or sequences) in such a way as to place expression of the sequence under the influence or control of the regulatory sequence. If the two DNA sequences are a coding sequence and a promoter region sequence linked to the 5' end of the coding sequence, they are operably linked if induction of promoter function results in the transcription of mRNA

encoding the desired protein and if the nature of the linkage between the two DNA sequences does not (1) result in the introduction of a frame-shift mutation, (2) interfere with the ability of the expression regulatory sequences to direct the expression of the protein, antisense RNA, or (3) interfere with the ability of the DNA template to be transcribed. Thus, a promoter region would be operably linked to a DNA sequence if the promoter was capable of effecting transcription of that DNA sequence.

The precise nature of the regulatory regions needed for gene expression can vary between species or cell types, but shall in general include, as necessary, 5' non-transcribing and 5' non-translating (non-coding) sequences involved with initiation of transcription and translation respectively, such as the TATA box, capping sequence, CAAT sequence, and the like, with those elements necessary for the promoter sequence being provided by the promoters of the invention. Such transcriptional control sequences can also include enhancer sequences or upstream activator sequences, as desired.

The vectors of the invention can further comprise other operably linked regulatory elements such as DNA elements which confer antibiotic resistance, or origins of replication for maintenance of the vector in one or more host cells.

In another embodiment, especially for maintenance of the vectors of the invention in prokaryotic cells, or in yeast *S. cerevisiae* cells, the introduced sequence is incorporated into a plasmid or viral vector capable of autonomous replication in the recipient host. Any of a wide variety of vectors can be employed for this purpose. In *Bacillus* hosts, integration of the desired DNA can be necessary.

Expression of a protein in eukaryotic hosts such as a human cell requires the use of regulatory regions functional in such hosts. A wide variety of transcriptional and translational regulatory sequences can be employed, depending upon the nature of the host. Preferably, these regulatory signals are associated in their native state with a particular gene which is capable of a high level of expression in the specific host cell, such as a specific human tissue type. In eukaryotes, where transcription is not linked to translation, such control regions may or may not provide an initiator methionine (AUG) codon, depending on whether the cloned sequence contains such a methionine. Such regions will, in general, include a promoter region sufficient to direct the initiation of RNA synthesis in the host cell.

If desired, the non-transcribed and/or non-translated regions 3' to the sequence coding for the huntingtin protein can be obtained by the above-described cloning methods. The 3'-non-transcribed region of the native human *huntingtin* gene can be retained for its transcriptional termination regulatory sequence elements, or for those elements which direct polyadenylation in eukaryotic cells. Where the native expression control sequences signals do not function satisfactorily in a host cell, then sequences functional in the host cell can be substituted.

It may be desired to construct a fusion product that contains a partial coding sequence (usually at the amino terminal end) of a first protein or small peptide and a second coding sequence (partial or complete) of the huntingtin protein at the carboxyl end. The coding sequence of the first protein can, for example, function as a signal sequence for secretion of the huntingtin protein from the host cell. Such first protein can also provide for tissue targeting or localization of the huntingtin protein if it is to be made in one cell type in a multicellular organism and delivered to another cell type in the same organism. Such fusion protein sequences can be designed with or without specific protease sites such that a desired peptide sequence is amenable to subsequent removal.

The expressed huntingtin protein can be isolated and purified from the medium of the host in accordance with conventional conditions, such as extraction, precipitation, chromatography, affinity chromatography, electrophoresis, or the like. For example, affinity purification with anti-huntingtin antibody can be used. A protein having the amino acid sequence shown in Figure 3 can be made, or a shortened peptide of this sequence can be made, and used to raise antibodies using methods well known in the art. These antibodies can be used to affinity purify or quantitate huntingtin protein from any desired source.

If it is necessary to extract huntingtin protein from the intracellular regions of the host cells, the host cells can be collected by centrifugation, or with suitable buffers, lysed, and the protein isolated by column chromatography, for example, on DEAE-cellulose, phosphocellulose, polyribocytidylic acid-agarose, hydroxyapatite or by electrophoresis or immunoprecipitation.

II. Use Of Huntingtin For Diagnostic And Treatment Purposes

It is to be understood that although the following discussion is specifically directed to human patients, the teachings are also applicable to any animal that expresses huntingtin and in which alteration of huntingtin, especially the amplification of CAG repeat copy number, leads to a defect in huntingtin gene (structure or function) or huntingtin protein (structure or function or expression), such that clinical manifestations such as those seen in Huntington's disease patients are found.

It is also to be understood that the methods referred to herein are applicable to any patient suspected of developing/having Huntington's disease, whether such condition is manifest at a young age or at a more advanced age in the patient's life. It is also to be understood that the term "patient" does not imply that symptoms are present, and patient includes any individual it is desired to examine or treat using the methods of the invention.

The diagnostic and screening methods of the invention are especially useful for a patient suspected of being at risk for developing Huntington's disease based on family history, or a patient in which it is desired to diagnose or eliminate the presence of the Huntington's disease condition as a causative agent behind a patient's symptoms.

It is to be understood that to the extent that a patient's symptoms arise due to the alteration of the CAG repeat copy numbers in the *huntingtin* gene, even without a diagnosis of Huntington's disease, the methods of the invention can identify the same as the underlying basis for such condition.

According to the invention, presymptomatic screening of an individual in need of such screening for their likelihood of developing Huntington's disease is now possible using DNA encoding the huntingtin gene of the invention, and specifically, DNA having the sequence of the normal human huntingtin gene. The screening method of the invention allows a presymptomatic diagnosis, including prenatal diagnosis, of the presence of an aberrant *huntingtin* gene in such individuals, and thus an opinion concerning the likelihood that such individual would develop or has developed Huntington's disease or symptoms thereof. This is especially valuable for the identification of carriers of altered huntingtin gene alleles where such alleles possess an increased number of CAG repeats in their huntingtin gene, for example, from individuals with a family history of Huntington's disease. Especially useful for the determination of the number of CAG repeats in the patient's *huntingtin* gene is the use of PCR to amplify such region or DNA blotting techniques.

For example, in the method of screening, a tissue sample would be taken from such individual, and screened for (1) the presence of the 'normal' human *huntingtin* gene, especially for the presence of a "normal" range of 11-34 CAG copies in such gene. The human *huntingtin* gene can be characterized based upon, for example, detection of restriction digestion patterns in 'normal' versus the patient's DNA, including RFLP analysis, using DNA probes prepared against the *huntingtin* sequence (or a functional fragment thereof) taught in the invention. Similarly, huntingtin mRNA can be characterized and compared to normal huntingtin mRNA (a) levels and/or (b) size as found in a human population not at risk of developing Huntington's disease using similar probes. Lastly, huntingtin protein can be (a) detected and/or (b) quantitated using a biological assay for huntingtin, for example, using an immunological assay and anti-huntingtin antibodies. When assaying huntingtin protein, the immunological assay is preferred for its speed. Methods of making antibody against the huntingtin are well known in the art.

An (1) aberrant *huntingtin* DNA size pattern, such as an aberrant *huntingtin* RFLP, and/or (2) aberrant huntingtin mRNA sizes or levels and/or (3) aberrant huntingtin protein levels would indicate that the patient has developed or is at risk for developing a huntingtin-associated symptom such as a symptom associated with Huntington's disease.

The screening and diagnostic methods of the invention do not require that the entire huntingtin DNA coding sequence be used for the probe. Rather, it is only necessary to use a fragment or length of nucleic acid that is sufficient to detect the presence of the huntingtin gene in a DNA preparation from a normal or affected individual, the absence of such gene, or an altered physical property of such gene (such as a change in electrophoretic migration pattern).

Prenatal diagnosis can be performed when desired, using any known method to obtain fetal cells, including amniocentesis, chorionic villous sampling (CVS), and fetoscopy. Prenatal chromosome analysis can be used to determine if the portion of chromosome 4 possessing the normal *huntingtin* gene is present in a heterozygous state, and PCR amplification or DNA blotting utilized for estimating the size of the CAG repeat in the *huntingtin* gene.

The huntingtin DNA can be synthesized, especially, the CAG repeat region can be amplified and, if desired, labeled with a radioactive or nonradioactive reporter group, using techniques known in the art (for example, see Eckstein, F., Ed., *Oligonucleotides and Analogues: A Practical Approach*, IRL Press at Oxford University Press, New York, 1992); and Kricka, L.J., Ed., *Nonisotopic DNA Probe Techniques*, Academic Press, San Diego, (1992)).

In one method of treating Huntington's disease in a patient in need of such treatment, functional *huntingtin* DNA is provided to the cells of such patient, preferably prior to such symptomatic state that indicates the death of many of the patient's neuronal cells which it is desired to target with the method of the invention. The replacement *huntingtin* DNA is provided in a manner and amount that permits the expression of the huntingtin protein provided by such gene, for a time and in a quantity sufficient to treat such patient. Many vector systems are known in the art to provide such delivery to human patients in need of a gene or protein missing from the

cell. For example, adenovirus or retrovirus systems can be used, especially modified retrovirus systems and especially herpes simplex virus systems. Such methods are provided for, in, for example, the teachings of Breakefield, X.A. *et al.*, *The New Biologist* 3:203-218 (1991); Huang, Q. *et al.*, *Experimental Neurology* 115:303-316 (1992), WO93/03743 and WO90/09441 each incorporated herein fully by reference. Methods of antisense strategies are known in the art (see, for example, *Antisense Strategies*, Baserga, R. *et al.*, Eds., Annals of the New York Academy of Sciences, volume 660, 1992).

In another method of treating Huntington's disease in a patient in need of such treatment, a gene encoding an expressible sequence that transcribes *huntingtin* antisense RNA is provided to the cells of such patient, preferably prior to such symptomatic state that indicates the death of many of the patient's neuronal cells which it is desired to target with the method of the invention. The replacement *huntingtin* antisense RNA gene is provided in a manner and amount that permits the expression of the antisense RNA provided by such gene, for a time and in a quantity sufficient to treat such patient, and especially in an amount to inhibit translation of the aberrant huntingtin mRNA that is being expressed in the cells of such patient. As above, many vector systems are known in the art to provide such delivery to human patients in need of a gene or protein which is altered in the patients' cells. For example, adenovirus or retrovirus systems can be used, especially modified retrovirus systems and especially herpes simplex virus systems. Such methods are provided for, in, for example, the teachings of Breakefield, X.A. *et al.*, *The New Biologist* 3:203-218 (1991); Huang, Q. *et al.*, *Experimental Neurology* 115:303-316 (1992), WO93/03743 and WO90/09441 each incorporated herein fully by reference.

Delivery of a DNA sequence encoding a functional huntingtin protein, such as the amino acid encoding sequence of Figure 4, will effectively replace the altered *huntingtin* gene of the invention, and inhibit, and/or stop and/or regress the symptoms that are the result of the interference to *huntingtin* gene expression due to an increased number of CAG repeats, such as 37 to 86 repeats in the *huntingtin* gene as compared to the 11-34 CAG repeats found in human populations not at risk for developing Huntington's disease.

Because Huntington's disease is characterized by a loss of neurons that is most severe in the caudate and putamen regions of the brain, the method of treatment of the invention is most effective when the replacement *huntingtin* gene is provided to the patient early in the course of the disease, prior to the loss of many neurons due to cell death. For that reason, presymptomatic screening methods according to the invention are important in identifying those individuals in need of treatment by the method of the invention, and such treatment preferably is provided while such individual is presymptomatic.

In a further method of treating Huntington's disease in a patient in need of such treatment such method provides an antagonist to the aberrant huntingtin protein in the cells of such patient.

Although the method is specifically described for DNA-DNA probes, it is to be understood that RNA possessing the same sequence information as the DNA of the invention can be used when desired.

For diagnostic assays, huntingtin antibodies are useful for quantitating and evaluating levels of huntingtin protein, and are especially useful in immunoassays and diagnostic kits.

In another embodiment, the present invention relates to an antibody having binding affinity to an huntingtin polypeptide, or a binding fragment thereof. In a preferred embodiment, the polypeptide has the amino acid sequence set forth in SEQ ID NO:6, or mutant or species variation thereof, or at least 7 contiguous amino acids thereof (preferably, at least 10, 15, 20, or 30 contiguous amino acids thereof). Those which bind selectively to huntingtin would be chosen for use in methods which could include, but should not be limited to, the analysis of altered huntingtin expression in tissue containing huntingtin.

The antibodies of the present invention include monoclonal and polyclonal antibodies, as well fragments of these antibodies. Antibody fragments which contain the idiotype of the molecule can be generated by known techniques. For example, such fragments include but are not limited to: the F(ab')₂ fragment; the Fab' fragments, and the Fab fragments.

Of special interest to the present invention are antibodies to huntingtin (or their functional derivatives) which are produced in humans, or are "humanized" (i.e. non-immunogenic in a human) by recombinant or other technology. Humanized antibodies may be produced, for example by replacing an immunogenic portion of an antibody with a corresponding, but non-immunogenic portion (i.e. chimeric antibodies) (Robinson, R.R. *et al.*, International Patent Publication PCT/US86/02269; Akira, K. *et al.*, European Patent Application 184,187; Taniguchi, M., European Patent Application 171,496; Morrison, S.L. *et al.*, European Patent Application 173,494; Neuberger, M.S. *et al.*, PCT Application WO 86/01533; Cabilly, S. *et al.*, European Patent Application 125,023; Better, M. *et al.*, *Science* 240:1041-1043 (1988); Liu, A.Y. *et al.*, *Proc. Natl. Acad. Sci. USA* 84:3439-3443 (1987); Liu, A.Y. *et al.*, *J. Immunol.* 139:3521-3526 (1987); Sun, L.K. *et al.*, *Proc. Natl. Acad. Sci. USA* 84:214-218 (1987); Nishimura, Y. *et al.*, *Canc. Res.* 47:999-1005 (1987); Wood, C.R. *et al.*, *Nature* 314:446-449 (1985); Shaw *et al.*, *J. Natl. Cancer Inst.* 80:1553-1559 (1988). General reviews of "humanized" chimeric antibodies are provided by Morrison, S.L. (*Science*, 229:1202-1207 (1985)) and by Oi, V.T. *et al.*, *BioTechniques* 4:214 (1986)). Suitable "humanized" antibodies can be alternatively produced by CDR or CEA substitution

(Jones, P.T. *et al.*, *Nature* 321:552-525 (1986); Verhoeyan *et al.*, *Science* 239:1534 (1988); Beidler, C.B. *et al.*, *J. Immunol.* 141:4053-4060 (1988)).

In another embodiment, the present invention relates to a hybridoma which produces the above-described monoclonal antibody, or binding fragment thereof. A hybridoma is an immortalized cell line which is capable of secreting a specific monoclonal antibody.

In general, techniques for preparing monoclonal antibodies and hybridomas are well known in the art (Campbell, *"Monoclonal Antibody Technology: Laboratory Techniques in Biochemistry and Molecular Biology,"* Elsevier Science Publishers, Amsterdam, The Netherlands (1984); St. Groth *et al.*, *J. Immunol. Methods* 35:1-21 (1980)).

Any animal (mouse, rabbit, and the like) which is known to produce antibodies can be immunized with the selected polypeptide. Methods for immunization are well known in the art. Such methods include subcutaneous or interperitoneal injection of the polypeptide. One skilled in the art will recognize that the amount of polypeptide used for immunization will vary based on the animal which is immunized, the antigenicity of the polypeptide and the site of injection.

The polypeptide may be modified or administered in an adjuvant in order to increase the peptide antigenicity. Methods of increasing the antigenicity of a polypeptide are well known in the art. Such procedures include coupling the antigen with a heterologous protein (such as globulin or β -galactosidase) or through the inclusion of an adjuvant during immunization.

For monoclonal antibodies, spleen cells from the immunized animals are removed, fused with myeloma cells, and allowed to become monoclonal antibody producing hybridoma cells.

Any one of a number of methods well known in the art can be used to identify the hybridoma cell which produces an antibody with the desired characteristics. These include screening the hybridomas with an ELISA assay, western blot analysis, or radioimmunoassay (Lutz *et al.*, *Exp. Cell Res.* 175:109-124 (1988)).

Hybridomas secreting the desired antibodies are cloned and the class and subclass is determined using procedures known in the art (Campbell, *Monoclonal Antibody Technology: Laboratory Techniques in Biochemistry and Molecular Biology, supra* (1984)).

For polyclonal antibodies, antibody containing antisera is isolated from the immunized animal and is screened for the presence of antibodies with the desired specificity using one of the above-described procedures.

In another embodiment of the present invention, the above-described antibodies are detectably labeled. Antibodies can be detectably labeled through the use of radioisotopes, affinity labels (such as biotin, avidin, and the like), enzymatic labels (such as horse radish peroxidase, alkaline phosphatase, and the like) fluorescent labels (such as FITC or rhodamine, and the like), paramagnetic atoms, and the like. Procedures for accomplishing such labeling are well-known in the art, for example, see (Sternberger *et al.*, *J. Histochem. Cytochem.* 18:315 (1970); Bayer *et al.*, *Meth. Enzym.* 62:308 (1979); Engval *et al.*, *Immunol.* 109:129 (1972); Goding, *J. Immunol. Meth.* 13:215 (1976)). The labeled antibodies of the present invention can be used for *in vitro*, *in vivo*, and *in situ* assays to identify cells or tissues which express a specific peptide.

In another embodiment of the present invention the above-described antibodies are immobilized on a solid support. Examples of such solid supports include plastics such as polycarbonate, complex carbohydrates such as agarose and sepharose, acrylic resins and such as polyacrylamide and latex beads. Techniques for coupling antibodies to such solid supports are well known in the art (Weir *et al.*, *"Handbook of Experimental Immunology"* 4th Ed., Blackwell Scientific Publications, Oxford, England, Chapter 10 (1986); Jacoby *et al.*, *Meth. Enzym.* 34 Academic Press, N.Y. (1974)). The immobilized antibodies of the present invention can be used for *in vitro*, *in vivo*, and *in situ* assays as well as in immunochromatography.

Furthermore, one skilled in the art can readily adapt currently available procedures, as well as the techniques, methods and kits disclosed above with regard to antibodies, to generate peptides capable of binding to a specific peptide sequence in order to generate rationally designed antipeptide peptides, for example see Hurby *et al.*, "Application of Synthetic Peptides: Antisense Peptides", In *Synthetic Peptides, A User's Guide*, W.H. Freeman, NY, pp. 289-307 (1992), and Kaspczak *et al.*, *Biochemistry* 28:9230-8 (1989).

Anti-peptide peptides can be generated in one of two fashions. First, the anti-peptide peptides can be generated by replacing the basic amino acid residues found in the huntingtin peptide sequence with acidic residues, while maintaining hydrophobic and uncharged polar groups. For example, lysine, arginine, and/or histidine residues are replaced with aspartic acid or glutamic acid and glutamic acid residues are replaced by lysine, arginine or histidine.

The manner and method of carrying out the present invention can be more fully understood by those of skill by reference to the following examples, which examples are not intended in any manner to limit the scope of the present invention or of the claims directed thereto.

Examples

The gene causing Huntington's disease has been mapped in 4p16.3 but has previously eluded identification. The invention uses haplotype analysis of linkage disequilibrium to spotlight a small segment of 4p16.3 as the likely location of the defect. A new gene, *huntingtin* (*IT15*), isolated using cloned "trapped" exons from a cosmid contig of the target area contains a polymorphic trinucleotide repeat that is expanded and unstable on HD chromosomes. A (CAG)_n repeat longer than the normal range of about 11 to about 34 copies was observed on HD chromosomes from all 75 disease families examined, comprising a wide range of ethnic backgrounds and 4p16.3 haplotypes. The (CAG)_n repeat, which varies from 37 to at least 86 copies on HD chromosomes appears to be located within the coding sequence of a predicted about 348 kDa protein that is widely expressed but unrelated to any known gene. Thus, the Huntington's disease mutation involves an unstable DNA segment, similar to those described in fragile X syndrome and myotonic dystrophy, acting in the context of a novel 4p16.3 gene to produce a dominant phenotype.

The following protocols and experimental details are referenced in the examples that follow.

HD Cell Lines. Lymphoblast cell lines from HD families of varied ethnic backgrounds used for genetic linkage and disequilibrium studies (Conneally *et al.*, *Genomics* 5:304-308 (1989); MacDonald *et al.*, *Nature Genet.* 1:99-103 (1992)) have been established (Anderson and Gusella, *In Vitro* 20:856-858 (1984)) in the Molecular Neurogenetics Unit, Massachusetts General Hospital, over the past 13 years. The Venezuelan HD pedigree is an extended kindred of over 10,000 members in which all affected individuals have inherited the HD gene from a common founder (Gusella *et al.*, *Nature* 306:234-238 (1983); Gusella *et al.*, *Science* 225:1320-1326 (1984); Wexler *et al.*, *Nature* 326:194-197 (1987)).

DNA/RNA Blotting. DNA was prepared from cultured cells and DNA blots prepared and hybridized as described (Gusella *et al.*, *Proc. Natl. Acad. Sci. USA* 76:5239-5243 (1979); Gusella *et al.*, *Nature* 306:234-238 (1983)). RNA was prepared and Northern blotting performed as described in Taylor *et al.*, *Nature Genet.* 3:223-227 (1992).

Construction of Cosmid Contig. The initial construction of the cosmid contig was by chromosome walking from cosmids L19 and BJ56 (Allitto *et al.*, *Genomics* 9:104-112 (1991); Lin *et al.*, *Somat. Cell Mol. Genet.* 17:481-488 (1991)). Two libraries were employed, a collection of Alu-positive cosmids from the reduced cell hybrid H39-8C10 (Whaley *et al.*, *Som. Cell Mol. Genet.* 17:83-91 (1991)) and an arrayed flow-sorted chromosome 4 cosmid library (NM87545) provided by the Los Alamos National Laboratory. Walking was accomplished by hybridization of whole cosmid DNA, using suppression of repetitive and vector sequences, to robot-generated high density filter grids (Nizetic, D. *et al.*, *Proc. Natl. Acad. Sci. USA* 88:3233-3237 (1991); Lehrach, H. *et al.*, in *Genome Analysis: Genetic and Physical Mapping, Volume 1*, Davies, K.E. *et al.*, Ed., Cold Spring Harbor Laboratory Press, 1991, pp. 39-81). Cosmids L1C2, L69F7, L228B6 and L83D3 were first identified by hybridization of YAC clone YGA2 to the same arrayed library (Bates *et al.*, *Nature Genet.* 1:180-187 (1992); Baxendale *et al.*, *Nucleic Acids Res.* 19:6651 (1991)). HD cosmid GUS72-2130 was isolated by standard screening of a GUS72 cosmid library using a single-copy probe. Cosmid overlaps were confirmed by a combination of clone-to-clone and clone-to-genomic hybridizations, single-copy probe hybridizations and restriction mapping.

cDNA Isolation and Characterization. Exon probes were isolated and cloned as described (Buckler *et al.*, *Proc. Natl. Acad. Sci. USA* 88:4005-4009 (1991)). Exon probes and cDNAs were used to screen human lambdaZAPII cDNA libraries constructed from adult frontal cortex, fetal brain, adenovirus transformed retinal cell line RCA, and liver RNA. cDNA clones, PCR products and trapped exons were sequenced as described (Sanger *et al.*, *Proc. Natl. Acad. Sci. USA* 74:5463-5467 (1977)). Direct cosmid sequencing was performed as described (McClatchey *et al.*, *Hum. Mol. Genet.* 1:521-527 (1992)). Database searches were performed using the BLAST network service of National Center for Biotechnology Information (Altschul *et al.*, *J. Mol. Biol.* 215:403-410 (1990)).

PCR Assay of the (CAG)_n Repeat. Genomic primers (SEQ ID NO:3 and SEQ ID NO:4) flanking the (CAG)_n repeat are:

5' ATG AAG GCC TTC GAG TCC CTC AAG TCC TTC 3'

and

5' AAA CTC ACG GTC GGT GCA GCG GCT CCT CAG 3'.

PCR amplification was performed in a reaction volume of 25 µl using 50 ng of genomic DNA, 5 µg of each primer, 10 mM Tris, pH 8.3, 50 mM KCl, 2 mM MgCl₂, 200 µM dNTPs, 10% DMSO, 0.1 unit Perfectmatch (Stra-

tagene), 2.5 μ Ci 32 P-dCTP (Amersham) and 1.25 units Taq polymerase (Boehringer Mannheim). After heating to 94°C for 1.5 minutes, the reaction mix was cycled according to the following program: 40 X [1'@94°C; 1'@60°C; 2'@72°C]. 5 μ l of each PCR reaction was diluted with an equal volume of 95 % formamide loading dye and heat denatured for 2 min. at 95°C. The products were resolved on 5 % denaturing polyacrylamide gels. The PCR product from this reaction using cosmid L191F1 (CAG₁₈) as template was 247 bp. Allele sizes were estimated relative to a DNA sequencing ladder, the PCR products from sequenced cosmids, and the invariant background bands often present on the gel. Estimates of allelic variation were obtained by typing unrelated individuals of largely Western European ancestry, and normal parents of affected HD individuals from various pedigrees.

Typing of HD and normal chromosomes in Examples 5-8. HD chromosomes were derived from symptomatic individuals and "at risk" individuals known to be gene carriers by linkage marker analysis. All HD chromosomes were from members of well-characterized HD families of varied ethnic backgrounds used previously for genetic linkage and disequilibrium studies (MacDonald, M.E., *et al.*, *Nature Genet.* 1:99-103 (1992); Conneally, P.M., *et al.*, *Genomics* 5:304-308 (1989)). Three of the 150 families used were large pedigrees, each descended from a single founder. The large Venezuelan HD pedigree is an extended kindred of over 13,000 members from which we typed 75 HD chromosomes (Gusella, J.F., *et al.*, *Nature* 306:234-238 (1983); Wexler, N.S., *et al.*, *Nature* 326:194-197 (1987)). Two other large families that have been described previously as Family Z and Family D, provided 25 and 35 HD chromosomes, respectively (Folstein, S.E., *et al.*, *Science* 229:776-779 (1985)). Normal chromosomes were taken from married-ins in the HD families and from unrelated normal individuals from non-HD families. The DNA tested for all individuals except four was prepared from lymphoblastoid cell lines or fresh blood (Gusella, J.F., *et al.*, *Nature* 306:234-238 (1983); Anderson and Gusella, *In Vitro* 20:856-858 (1984)). In the exceptional cases, DNA was prepared from frozen cerebellum. No difference in the characteristics of the PCR products were observed between lymphoblastoid, fresh blood, or brain DNAs. For five members of the Venezuelan pedigree aged 24-30, we also prepared DNA by extracting pelleted sperm from semen samples. The length of the HD gene (CAG)_n repeat for all DNAs was assessed using polymerase chain reaction amplification.

Statistical analysis as set forth in Examples 5-8. Associations between repeat lengths and onset age were assessed by Pearson correlation coefficient and by multivariate regression to assess higher order associations. Comparisons of the distributions of repeat length for all HD chromosomes and those for individual families were made by analysis of variance and t-test contrasts between groups. The 95 % confidence bands were computed around the regression line utilizing the general linear models procedure of SAS (SAS Institute Inc., SAS/STAT User's Guide, Version 6, Fourth Edition, Volume 2 (SAS Institute Inc., Cary, N.C., pp. 846, 1989)).

Example 1

Application of Exon Amplification to Obtain Trapped Cloned Exons

The HD candidate region defined by discrete recombination events in well-characterized families spans 2.2 Mb between D4S10 and D4S98 as shown in Figure 1. The 500 kb segment between D4S180 and D4S182 displays the strongest linkage disequilibrium with HD, with about 1/3 of disease chromosomes sharing a common haplotype, anchored by multi-allele polymorphisms at D4S127 and D4S95 (MacDonald *et al.*, *Nature Genet.* 1:99-103 (1992)). Sixty-four overlapping cosmids spanning about 480 kb from D4S180 to a location between D4S95 and D4S182 have been isolated by a combination of information from YAC (Baxendale *et al.*, *Nucleic Acids Res.* 19:6651 (1991)) and cosmid probe hybridization to high density filter grids of a chromosome 4 specific library, as well as additional libraries covering this region. Sixteen of these cosmids providing the complete contig are shown in Figure 1. We have previously used exon amplification to identify ADDA, the α -adducin locus, IT10C3, a novel putative transporter gene, and IT11, a novel G protein-coupled receptor kinase gene in the region distal to D4S127 (Figure 1).

We have now applied the exon amplification technique to cosmids from the region of the contig proximal to D4S127. This procedure produces "trapped" exon clones, which can represent single exons, or multiple exons spliced together and is an efficient method of obtaining probes for screening cDNA libraries. Individual cosmids were processed, yielding 9 exon clones in the region from cosmids L134B9 to L181B10.

Two non-overlapping cDNAs were initially isolated using exon probes. IT15A was obtained by screening a transformed adult retinal cell cDNA library with exon clone DL118F5-U. IT16A was isolated by screening an adult frontal cortex cDNA library with a pool of three exon clones, DL83D3-8, DL83D3-1, and DL228B6-3. By Northern blot analysis, we discovered that IT15A and IT16A are in fact different portions of the same large approximately 10-11 kb transcript. Figure 2 shows an example of a Northern blot containing RNA from lymphoblastoid cell lines representing a normal individual and 2 independent homozygotes for HD chromosomes

of different haplotypes. The same approximately 10-11 kb transcript was also detected in RNA from a variety of human tissues (liver, spleen, kidney, muscle and various regions of adult brain).

IT15A and IT16A were used to "walk" in a number of human tissue cDNA libraries in order to obtain the full-length transcript. Figure 3 shows a representation of 5 cDNA clones which define the IT15 transcript, under a schematic of the composite sequence derived as described in the legend. Figure 3 also displays the locations on the composite sequence of the 9 trapped exon clones.

The composite sequence of IT15, containing the entire predicted coding sequence, spans 10,366 bases including a tail of 18 A's as shown in Figure 4. An open reading frame of 9,432 bases begins with a potential initiator methionine codon at base 316, located in the context of an optimal translation initiation sequence. An in-frame stop codon is located 240 bases upstream from this site. The protein product of IT15 is predicted to be a 348 kDa protein containing 3,144 amino acids. Although the first Met codon in the long open reading frame has been chosen as the probably initiator codon, we cannot exclude that translation does not actually begin at a more 3' Met codon, producing a smaller protein.

Example 2

Polymorphic Variation of the (CAG)_n Trinucleotide Repeat

Near its 5' end, the IT15 sequence contains 21 copies of the triplet CAG, encoding glutamine (Figure 5). When this sequence was compared with genomic sequences that are known to surround simple sequence repeats (SSRs) in 4p16.3, it was found that normal cosmid L191F1 had 18 copies of the triplet indicating that the (CAG)_n repeat is polymorphic (Figure 5). Primers from the genomic sequence flanking the repeat were chosen to establish a PCR assay for this variation. In the normal population, this SSR polymorphism displays at least 17 discrete alleles (Table 1) ranging from about 11 to about 34 repeat units. Ninety-eight percent of the 173 normal chromosomes tested contained repeat lengths between 11 and 24 repeats. Two chromosomes were detected in the 25-30 repeat range and 2 normal chromosomes had 33 and 34 repeats respectively. The overall heterozygosity on normal chromosome was 80%. Based on sequence analysis of three clones, it appears that the variation is based entirely on the (CAG)_n, but the potential for variation of the smaller downstream (CCG)₇ which is also included in the PCR product, is also present.

Example 3

Instability of the Trinucleotide Repeat on HD chromosomes

Sequence analysis of cosmid GUS72-2130, derived from a chromosome with the major HD haplotype (see below), revealed 48 copies of the trinucleotide repeat, far greater than the largest normal allele (Figure 5). When the PCR assay was applied to HD chromosomes, a pattern strikingly different from the normal variation was observed. HD heterozygotes contained one discrete allelic product in the normal size range, and one PCR product of much larger size, suggesting that the (CAG)_n repeat on HD chromosomes is expanded relative to normal chromosomes.

Figure 6 shows the patterns observed when the PCR assay was performed on lymphoblast DNA from a selected nuclear family in a large Venezuelan HD kindred. In this family, DNA marker analysis has shown previously that the HD chromosome was transmitted from the father (lane 2) to seven children (lanes 3, 5, 6, 7, 8, 10 and 11). The three normal chromosomes present in this mating yielded a PCR product in the normal size range (AN1, AN2, AN3) that was inherited in a Mendelian fashion. The HD chromosome in the father yielded a diffuse, "fuzzy"-appearing PCR product slightly smaller than the 48 repeat product of the non-Venezuelan HD cosmid. Except for the DNA in lane 5 which did not PCR amplify and in lane 11 which displayed only a single normal allele, each of the affected children's DNAs yielded a fuzzy PCR product of a different size (AE), indicating instability of the HD chromosome (CAG)_n repeat. Lane 6 contained an HD-specific product slightly smaller than or equal to that of the father's DNA. Lanes 3, 7, 10 and 8, respectively, contained HD-specific PCR products of progressively larger size. The absence of an HD-specific PCR product in lane 11 suggested that this child's DNA possessed a (CAG)_n repeat that was too long to amplify efficiently. This was verified by Southern blot analysis in which the expanded HD allele was easily detected and estimated to contain up to 100 copies of the repeat. Notably, this child had juvenile onset of HD at the very early age of 2 years. The onset of HD in the father was in his early 40s, typical of most adult HD patients in this population. The onset ages of children represented by lanes 3, 7, 10 and 8 were 26, 25, 14 and 11 years, respectively, suggesting a rough correlation between age at onset of HD and the length of the (CAG)_n repeat on the HD chromosome. In keeping with this trend, the offspring represented in lane 6 with the fewest repeats remained asymptomatic

when last examined at age of 30.

Figure 7 shows PCR analysis for a second sibship from the Venezuelan pedigree in which both parents are *HD* heterozygotes carrying the same *HD* chromosome based on DNA marker studies. Several of the offspring are *HD* homozygotes (lanes 6+7, 10+11, 13+14, 17+18, 23+24) as reported previously (Wexler *et al.*, *Nature* 326:194-197 (1987)). Each parent's DNA contained one allele in the normal range (AN1, AN2) which was transmitted in a Mendelian fashion. The *HD*-specific products (AE) from the DNA of both parents and children were all much larger than the normal allelic products and also showed extensive variation in mean size. A neurologic diagnosis for the offspring in this pedigree was not provided to maintain the blind status of investigators involved in the ongoing Venezuela *HD* project, although age of onset again appears to parallel repeat length. Paired samples under many of the individual symbols represent independent lymphoblast lines initiated at least one year apart. The variance between paired samples was not as great as between the different individuals, suggesting that the major differences in size of the PCR products resulted from meiotic transmission. Of special note is the result obtained in lanes 13 and 14. This *HD* homozygote's DNA yielded one PCR product larger and one smaller than the *HD*-specific PCR products of both parents.

To date, we have tested 75 independent *HD* families, representing all different reported in MacDonald *et al.*, *Nature Genet.* 1:99-103 (1992)) and a wide range of ethnic backgrounds. In all 75 cases, a PCR product larger than the normal size range was produced from the *HD* chromosome. The sizes of the *HD*-specific products ranged from 42 repeat copies to more than 66 copies, with a few individuals failing to yield a product because of the extreme length of the repeat. In these cases, Southern blot analysis revealed an increase in the length of an *EcoRI* fragment with the largest allele approximating 100 copies of the repeat. Figure 8 shows the variation detected in members of an American family of Irish ancestry in which the major *HD* haplotype is segregating. Cosmid GUS72-2130 was cloned from the *HD* homozygous individual whose DNA was amplified in lane 2. As was observed in the Venezuelan *HD* pedigree (Figures 6 and 7), which segregates the disorder with a different 4p16.3 haplotype, the *HD*-specific PCR products for this family display considerable size variation.

Example 4

New Mutations to *HD*

The mutation rate in *HD* has been reported to be very low. To test whether the expansion of the (CAG)_n repeat is the mechanism by which new *HD* mutations occur, two pedigrees with sporadic cases of *HD* have been examined in which intensive searching failed to reveal a family history of the disorder. In these cases, pedigree information sufficient to identify the same chromosomes in both the affected individual and unaffected relatives was gathered. Figures 9 and 10 show the results of PCR analysis of the (CAG)_n repeat in these families. The chromosomes in each family were assigned an arbitrary number based on typing for a large number of RFLP and SSR markers in 4p16.3 defining distinct haplotypes and the presumed *HD* chromosome is starred.

In family #1, *HD* first appeared in individual II-3 who transmitted the disorder to III-1 along with chromosome 3*. This same chromosome was present in II-2, an elderly unaffected individual. PCR analysis revealed that chromosome 3* from II-2 produced a PCR product at the extreme high end of the normal range (about 36 CAG copies). However, the (CAG)_n repeat on the same chromosome in II-3 and III-1 had undergone sequential expansions to about 44 and about 46 copies, respectively. A similar result was obtained in Family #2, where the presumed *HD* mutant III-2 had a considerably expanded repeat relative to the same chromosome in II-1 and III-1 (about 49 vs. about 33 CAG copies). In both family #1 and family #2, the ultimate *HD* chromosome displays the marker haplotype characteristic of 1/3 of all *HD* chromosomes, suggesting that this haplotype may be predisposed to undergoing repeat expansion.

Discussion

The discovery of an expanded, unstable trinucleotide repeat on *HD* chromosomes within the *IT15* gene is the basis for utilizing this gene as the *HD* gene of the invention. These results are consistent with the interpretation that *HD* constitutes the latest example of a mutational mechanism that may prove quite common in human genetic disease. Elongation of a trinucleotide repeat sequence has been implicated previously as the cause of three quite different human disorders, the fragile X syndrome, myotonic dystrophy and spino-bulbar muscular atrophy. The initial observations of repeat expansion in *HD* indicate that this phenomenon shares features in common with each of these disorders.

In the fragile X syndrome, expression of a constellation of symptoms that includes mental retardation and

a fragile site at Xq27.3 is associated with expansion of a (CGG)_n repeat thought to be in the 5' untranslated region of the *FMR1* gene (Fu *et al.*, *Cell* 67:1047-1058 (1991); Kremer *et al.*, *Science* 252:1711-1714 (1991); Verkerk *et al.*, *Cell* 65:904-914 (1991)). In myotonic dystrophy, a dominant disorder involving muscle weakness with myotonia that typically present in early adulthood, the unstable trinucleotide repeat, (CTG)_n, is located in the 3' untranslated region of the myotonic protein kinase gene (Aslanidis *et al.*, *Nature* 355:548-551 (1992); Brook *et al.*, *Cell* 68:799-808 (1992); Buxton *et al.*, *Nature* 355:547-548 (1992); Fu *et al.*, *Science* 255:1256-1259 (1992); Harley *et al.*, *Lancet* 339:1125-1128 (1992); Mahadevan *et al.*, *Science* 255:1253-1255 (1992)). The unstable (CAG)_n repeat in HD may be within the coding sequence of the IT15 gene, a feature shared with spino-bulbar muscular atrophy, an X-linked recessive adult-onset disorder of the motor neurons caused by expansion of a (CAG)_n repeat in the coding sequence of the androgen receptor gene (LaSpada *et al.*, *Nature* 352:77-79 (1991)). The repeat length in both the fragile X syndrome and myotonic dystrophy tends to increase in successive generations, sometimes quite dramatically. Occasionally, decreases in the average repeat length are observed (Fu *et al.*, *Science* 255:1256-1259 (1992); Yu *et al.*, *Am. J. Hum. Genet.* 50:968-980 (1992); Bruner *et al.*, *N. Engl. J. Med.* 476-480 (1993)). The HD trinucleotide repeat is also unstable, usually expanding when transmitted to the next generation, but contracting on occasion. In HD, as in the other disorders, change in copy number occurs in the absence of recombination. Compared with the fragile X syndrome, myotonic dystrophy, and HD, the instability of the disease allele in spino-bulbar muscular atrophy is more limited, and dramatic expansions of repeat length have not been seen (Biancalana *et al.*, *Hum. Mol. Genet.* 1:255-258 (1992)).

Expansion of the repeat length in myotonic dystrophy is associated with a particular chromosomal haplotype, suggesting the existence of a primordial predisposing mutation (Harley *et al.*, *Am. J. Hum. Genet.* 49:68-75 (1991); Harley *et al.*, *Nature* 355:545-546 (1992); Ashizawa, *Lancet* 338:642-643 (1991); and Epstein (1991)). In the fragile X syndrome, there may be a limited number of ancestral mutations that predispose to increases in trinucleotide repeat number (Richards *et al.*, *Nature Genet.* 1:257-260 (1992); Oudet *et al.*, *Am. J. Hum. Genet.* 52:297-304 (1993)). The linkage disequilibrium analysis used to identify IT15 indicates that there are several haplotypes associated with HD, but that at least 1/3 of HD chromosomes are ancestrally related (MacDonald *et al.*, *Nature Genet.* 1:99-103 (1992)). These data, combined with the reported low rate of new mutation to HD (Harper, *J. Med. Genet.* 89:365-376 (1992)), suggest that expansion of the trinucleotide repeat may only occur on select chromosomes. The analysis of two families presented herein, in which new mutation was supposed to have occurred, is consistent with the view that there may be particular normal chromosomes that have the capacity to undergo expansion of the repeat into the HD range. In each of these families, a chromosome with a (CAG)_n repeat length in the upper end of the normal range was segregating on a chromosome whose 4p16.3 haplotype matched the most common haplotype seen on HD chromosomes and the clinical appearance of HD in these two cases was associated with expansion of the trinucleotide repeat.

The recent application of haplotype analysis to explore the linkage disequilibrium on HD chromosomes pointed to a portion of a 2.2 Mb candidate region defined by the majority of recombination events described in HD pedigrees (MacDonald *et al.*, *Nature Genet.* 1:99-103 (1992)). Previously, the search for the gene was confounded by three matings in which the genetic inheritance pattern was inconsistent with the remainder of the family (MacDonald *et al.*, *Neuron* 3:183-190 (1989b); Prichard *et al.*, *Am. J. Hum. Genet.* 50:1218-1230 (1992)). These matings produced apparently affected HD individuals despite the inheritance of only normal alleles for markers throughout 4p16.3, effectively excluding inheritance of the HD chromosome present in the rest of the pedigree. Using PCR assay disclosed above, each of these families was tested and it was determined that like other HD kindreds, an expanded allele segregates with HD in affected individuals of all three pedigrees. However, an expanded allele was not present in those specific individuals with the inconsistent 4p16.3 genotypes. Instead, these individuals displayed the normal alleles expected based on analysis of other markers in 4p16.3. It is conceivable that these inconsistent individuals do not, in fact, have HD, but some other disorder. Alternatively, they might represent genetic mosaics in which the HD allele is more heavily represented and/or more expanded in brain tissue than in the lymphoblast DNA used for genotyping.

The capacity to monitor directly the size of the trinucleotide repeat in individuals "at risk" for HD provides significant advantages over current methods, eliminating the need for complicated linkage analyses, facilitating genetic counseling, and extending the applicability of presymptomatic and prenatal diagnosis to "at risk" individuals with no living affected relatives. However, it is of the utmost importance that the current internationally accepted guidelines and counseling protocols for testing those "at risk" continue to be observed, and that samples from unaffected relatives should not be tested inadvertently or without full consent. In the series of patients examined in this study, there is an apparent correlation between repeat length and age of onset of the disease, reminiscent of that reported in myotonic dystrophy (Harley *et al.*, *Lancet* 339:1125-1128 (1992); Tsilfidis *et al.*, *Nature Genet.* 1:192-195 (1992)). The largest HD trinucleotide repeat segments were found in juvenile onset cases, where there is a known preponderance of male transmission (Merrit *et al.*, *Excerpta Medica*, Amsterdam, pp. 645-650 (1969)).

The expression of fragile X syndrome is associated with direct inactivation of the *FMR1* gene (Pierretti *et al.*, *Cell* 66:817-822 (1991); DeBoule *et al.*, *Nature Genet.* 3:31-35 (1993)). The recessive inheritance pattern of spino-bulbar muscular atrophy suggests that in this disorder, an inactive gene product is produced. In myotonic dystrophy, the manner in which repeat expansion leads to the dominant disease phenotype is unknown. There are numerous possibilities for the mechanism of pathogenesis of the expanded trinucleotide repeat in HD. Without intending to be held to this theory, nevertheless notice can be taken that since Wolf-Hirschhorn patients hemizygous for 4p16.3 do not display features of HD, and IT15 mRNA is present in HD homozygotes, the expanded trinucleotide repeat does not cause simple inactivation of the gene containing it. The observation that the phenotype of HD is completely dominant, since homozygotes for the disease allele do not differ clinically from heterozygotes, has suggested that HD results from a gain of function mutation, in which either the mRNA product or the protein product of the disease allele would have some new property, or be expressed inappropriately (Wexler *et al.*, *Nature* 326:194-197 (1987); Myers *et al.*, *Am. J. Hum. Genet.* 45:615-618 (1989)). If the expanded trinucleotide repeat were translated, the consequences on the protein product would be dramatic, increasing the length of the poly-glutamine stretch near the N-terminus. It is possible, however, that despite the presence of an upstream Met codon, the normal translational start occurs 3' to the (CAG)_n repeat and there is no poly-glutamine stretch in the protein product. In this case, the repeat would be in the 5' untranslated region and might be expected to have its dominant effect at the mRNA level. The presence of an expanded repeat might directly alter regulation, localization, stability or translatability of the mRNA containing it, and could indirectly affect its counterpart from the normal allele in HD heterozygotes. Other conceivable scenarios are that the presence of an expanded repeat might alter the effective translation start site for the HD transcript, thereby truncating the protein, or alter the transcription start site for the IT15 gene, disrupting control of mRNA expression. Finally, although the repeat is located within the IT15 transcript, the possibility that it leads to HD by virtue of an action on the expression of an adjacent gene cannot be excluded.

Despite this final caveat, it is consistent with the above results and most likely that the trinucleotide repeat expansion causes HD by its effect, either at the mRNA or protein level, on the expression and/or structure of the protein product of the IT15 gene, which has been named huntingtin. Outside of the region of the triplet repeat, the IT15 DNA sequence detected no significant similarity to any previously reported gene in the GenBank database. Except for the stretches of glutamine and proline near the N-terminus, the amino acid sequence displayed no similarity to known proteins, providing no conspicuous clues to huntingtin's function. The poly-glutamine and poly-proline region near the N-terminus detect similarity with a large number of proteins which also contain long stretches of these amino acids. It is difficult to assess the significance of such similarities, although it is notable that many of these are DNA binding proteins and that huntingtin does have a single leucine zipper motif at residue 1,443. Huntingtin appears to be widely expressed, and yet cell death in HD is confined to specific neurons in particular regions of the brain.

TABLE 1. COMPARISON OF HD AND
NORMAL REPEAT SIZES

RANGE OF ALLELE SIZES (#REPEATS)	NORMAL CHROMOSOMES NUMBER AND FREQUENCY		HD CHROMOSOMES NUMBER AND FREQUENCY	
≥ 48	0	0	44	0.59
42-47	0	0	30	0.41
30-41	2	0.01	0	0
25-30	2	0.01	0	0
≤ 24	169	0.98	0	0
TOTAL	173	1.00	74	1.0

Example 5

Distribution of Trinucleotide Repeat Lengths on Normal and HD Chromosomes

The number of copies of the HD triplet repeat has been examined in a total of 425 HD chromosomes from 150 independent families and compared with the copy number of the HD triplet repeat of 545 normal chromosomes. The results are displayed in Figure 11. Two non-overlapping distributions of repeat length were observed, wherein the upper end of the normal range and the lower end of the HD range were separated by 3 repeat units. The normal chromosomes displayed 24 alleles producing PCR products ranging from 11 to 34 repeat units, with a median of 19 units (mean 19.71, s.d. 3.21). The HD chromosomes yielded 54 discrete PCR products corresponding to repeat lengths of 37 to 86 units, with a median of 45 units (mean 46.42, s.d. 6.68).

Of the HD chromosomes, 134 and 161 were known to be maternally or paternally-derived, respectively. To investigate whether the sex of the transmitting parent might influence the distribution of repeat lengths, these two sets of chromosomes were plotted separately in Figure 12. The maternally-derived chromosomes displayed repeat lengths ranging from 37 to 73 units, with a median of 44 (mean 44.93, s.d. 5.14). The paternally-derived chromosomes had 37 to 86 copies of the repeat unit, with a median of 48 units (mean 49.14, s.d. 8.27). However, a higher proportion of the paternally-derived HD chromosomes had repeat lengths greater than 55 units (16% vs. 2%), suggesting the possibility of a differential effect of paternal versus maternal transmission.

The data set used excluded chromosomes from a few clinically diagnosed individuals who have previously been shown not to have inherited the HD chromosome by DNA marker linkage studies (MacDonald, M.E., *et al.*, *Neuron* 3:183-190 (1989); Pritchard, C., *et al.*, *Am. J. Hum. Genet.* 50:1218-1230 (1992)). These individuals have repeat lengths well within the normal range. Their disease manifestations have not been explained, and they may represent phenocopies of HD. Regardless of the mechanism involved, the occurrence at low frequency of such individuals within known HD families must be considered if diagnostic conclusions are based solely on repeat length.

The control data set also excludes a number of chromosomes from phenotypically normal individuals who are related to "spontaneous" cases of HD or "new mutations". Chromosomes from these individuals who are not clinically affected and have no family history of the disorder cannot be designated as HD. However, these chromosomes cannot be classified as unambiguously normal because they are essentially the same chromo-

some as that of an affected relative, the diagnosed "spontaneous" HD proband, except with respect to repeat length. The lengths of repeat found on these ambiguous chromosomes (34-38 units) span the gap between the control and HD distributions, confounding a decision on the status of any individual with a repeat in the high normal to low HD range.

Example 6

Instability of the Trinucleotide Repeat

The data in Figure 11 combine repeat lengths from 150 different HD families representing many potentially independent origins of the defect. To examine the variation in repeat lengths on sets of HD chromosomes known to descend from a common founder, the data from three large HD kindreds (Gusella, J.F., *et al.*, *Nature* 306:234-238 (1983); Wexler, N.S., *et al.*, *Nature* 326:194-197 (1987); Folstein, S.E., *et al.*, *Science* 229:776-779 (1985)) with different 4p16.3 haplotypes (MacDonald, M.E., *et al.*, *Nature Genet.* 1:99-103 (1992)), typed for 75, 25 and 35 individuals, respectively, were separated. Despite the single origin of the founder HD chromosome within each pedigree, members of the separate pedigrees display a wide range of repeat lengths (Figure 13). This instability of the HD chromosome repeat is most prominent in members of a large Venezuelan HD kindred (panel A) in which the common HD ancestor has produced 10 generations of descendants, numbering over 13,000 individuals. The distribution of repeat lengths in this sampling of the Venezuelan pedigree (median 46, mean 48.26, s.d. 9.3) is not significantly different from that of the larger sample of HD chromosomes from all families. Panels B and C display results for two extended families in which HD was introduced more recently than in the Venezuelan kindred. These families have been reported to exhibit different age of onset distributions and varied phenotypic features of HD (Folstein, S.E., *et al.*, *Science* 229:776-779 (1985)). Both revealed extensive repeat length variation, with a median of 41 and 49 repeat units, respectively. The distribution of repeat lengths in the members of the family in Panel B was significantly different from the distribution of all HD chromosome repeat lengths ($p < 0.0001$), with a smaller mean of 42.04 repeat units (s.d. 2.82). The repeat distribution from HD chromosomes of Panel C was also significantly different from the total data set ($p < 0.004$), but with a higher mean of 49.80 (s.d. 5.86).

Example 7

Parental Source Effects on Repeat Length Variation

For 62 HD chromosomes in Figure 11, the length of the trinucleotide repeat also could be examined on the corresponding parental HD chromosome. In 20 of 25 maternal transmissions, and in 31 of 37 paternal transmissions, the repeat length was altered, indicating considerable instability. A similar phenomenon was not observed for normal chromosomes, where more than 500 meiotic transmissions revealed no changes in repeat length, although the very existence of such a large number of normal alleles suggests at least a low degree of instability.

Figure 14 shows the relationship between the repeat lengths on the HD chromosomes in the affected parent and corresponding progeny. For the 20 maternally-inherited chromosomes on which the repeat length was altered, 13 changes were increases in length and 7 were decreases. Both increases and decreases involved changes of less than 5 repeat units and the overall correlation between the mother's repeat length and that of her child was $r = 0.95$ ($p < 0.0001$). The average change in repeat length in the 25 maternal transmissions was an increase of 0.4 repeats.

On paternally-derived chromosomes, the 31 transmissions in which the repeat length changes comprised 26 length increases and 5 length decreases. Although the decreases in size were only slightly smaller than those observed on maternally-derived chromosomes, ranging from 1 to 3 repeat units, the increases were often dramatically larger. Thus, the correlation of the repeat length in the father with that of his offspring was only $r = 0.35$ ($p < 0.04$). The average change in the 37 paternal transmissions was an increase of 9 repeat units. The maximum length increase observed through paternal transmission was 41 repeat units, a near doubling of the parental repeat.

For both male and female transmissions, there was no correlation between the size of the parental repeat and either the magnitude or frequency of changes.

To determine whether the variation in the length of the repeat observed through male transmission of HD chromosomes is reflected in the male germ cells, we amplified the repeat from sperm DNA and from DNA of the corresponding lymphoblast from 5 HD gene carriers. The results, shown in Figure 15, reveal striking differences between the lymphoblast and sperm DNA for the HD chromosome repeat, but not for the repeat on

the normal chromosome. All the sperm donors are members of the Venezuelan HD family and range in age from 24 to 30 years. Individuals 1 and 2 are siblings with HD chromosome repeat lengths based on lymphoblast DNA of 45 and 52, respectively. Individuals 3 and 4 are also siblings, with HD repeat lengths of 46 and 49, respectively. Individual 5, from a different sibship than either of the other two pairs, has an HD repeat of 52 copies. In all 5 cases, the PCR amplification of sperm DNA and lymphoblast DNA yielded identical products from the normal chromosome. However, in comparison with lymphoblast DNA, the HD gene from sperm DNA yielded a diffuse array of products. In 3 of the 5 cases (2,4 and 5), the diffuse array spread to much larger allelic products than the corresponding lymphoblast product. Subject 2 showed the greatest range of expansion, with the sperm DNA product extending to over 80 repeat units. Interestingly, the 3 individuals displaying the greatest variation have the longest repeats and are currently symptomatic. The other two donors have shorter repeat lengths in the HD range, and remain at risk at this time.

The striking difference in the high repeat length range (>55) between HD chromosomes transmitted from the father and those transmitted from the mother indicated a potential parental source effect. When this was examined directly, the HD chromosome repeat length changed in about 85% of transmissions. Most changes involved a fluctuation of only a few repeat units, with larger increases occurring only in male transmissions. The greater size increases in male transmission appear to be caused by particular instability of the HD trinucleotide repeat during male gametogenesis, based on the amplification of the repeat from sperm DNA.

Example 8

Relationship between Repeat Length and Age of Onset

Increased repeat length might correlate with a reduced age of onset of HD. Accordingly, age of onset data was determined for 234 of the individuals represented in Figure 11. Figure 16 displays the repeat lengths found on the HD and normal chromosomes of these individuals relative to their age of onset. Indeed, age of onset is inversely correlated with the HD repeat length. A Pearson correlation coefficient of $r=-.75$, $p<0.0001$ was obtained assuming a linear relationship between age of onset and repeat length. When a polynomial function was used, a better fit was obtained ($R^2=0.61$, $F=121.45$), suggesting a higher order association between age of onset and repeat length.

There is considerable variation in the age of onset associated with any specific number of repeat units, particularly for trinucleotide repeats in the 37-52 unit zone (88% of HD chromosomes) where onset ranged from 15 to 75 years. In this range, a linear relationship between age of onset and repeat length provided as good a fit as a higher order relationship. The 95 % confidence interval surrounding the predicted regression line was estimated at ± 18 years. In the 37 to 52 unit range, the association of repeat length to onset age is only half as strong as in the overall distribution ($r=-0.40$, $p<0.0001$), indicating that much of the predictive power is contributed by repeats longer than 52 units. In this increased range, onset is likely to be very young and consequently not relevant to most persons seeking testing.

For the 178 cases in the 37-52 repeat unit range for which it was possible to subdivide the data set based on parental origin of the HD gene, multivariate regression analysis suggested a significant effect of parental origin on age of onset ($p<0.05$) independent of repeat length in this range. HD gene carriers from maternal transmissions had an average age of onset two years later than those from paternal transmissions.

In both univariate and multivariate analyses, no association between age of onset and the repeat length on the normal chromosome was detected, either in the total data set, or when it was subdivided into chromosomes of maternal or paternal origin.

All publications mentioned hereinabove are hereby incorporated in their entirety by reference.

While the foregoing invention has been described in some detail for purposes of clarity and understanding, it will be appreciated by one skilled in the art from a reading of this disclosure that various changes in form and detail can be made without departing from the true scope of the invention and appended claims.

SEQUENCE LISTING

5 (1) GENERAL INFORMATION:

(i) APPLICANT: THE GENERAL HOSPITAL CORPORATION
Fruit Street
Boston, Massachusetts 02114
United States of America

10 (ii) TITLE OF INVENTION: Huntingtin DNA, Protein And Uses Thereof

(iii) NUMBER OF SEQUENCES: 6

(iv) CORRESPONDENCE ADDRESS:

15 (A) KILBURN & STRODE
(B) 30 JOHN STREET
(C) LONDON
(D) GREAT BRITAIN
(E) WC1N 2DD

20 (v) COMPUTER READABLE FORM:

(A) MEDIUM TYPE: Floppy disk
(B) COMPUTER: IBM PC compatible
(C) OPERATING SYSTEM: PC-DOS/MS-DOS
(D) SOFTWARE: PatentIn Release #1.0, Version #1.25

25 (vi) CURRENT APPLICATION DATA:

(A) 7th March 1994

(vii) PRIOR APPLICATION DATA:

30 (A) APPLICATION NUMBER: 08/085,000
(B) FILING DATE: 01 JULY 1993

(vii) PRIOR APPLICATION DATA:

(A) APPLICATION NUMBER: 08/027,498
(B) FILING DATE: 05 MARCH 1993

35 (2) INFORMATION FOR SEQ ID NO:1:

(i) SEQUENCE CHARACTERISTICS:

40 (A) LENGTH: 20 base pairs
(B) TYPE: nucleic acid
(C) STRANDEDNESS: single
(D) TOPOLOGY: linear

(xi) SEQUENCE DESCRIPTION: SEQ ID NO:1:

GGCGGGAGAC CGCCATGGCG

20

45 (2) INFORMATION FOR SEQ ID NO:2:

(i) SEQUENCE CHARACTERISTICS:

50 (A) LENGTH: 17 base pairs
(B) TYPE: nucleic acid
(C) STRANDEDNESS: single
(D) TOPOLOGY: linear

(xi) SEQUENCE DESCRIPTION: SEQ ID NO:2:

AATACGACTC ACTATAG

17

55

- (i) SEQUENCE CHARACTERISTICS:
 (A) LENGTH: 30 base pairs
 (B) TYPE: nucleic acid
 (C) STRANDEDNESS: single
 (D) TOPOLOGY: linear

5

- (xi) SEQUENCE DESCRIPTION: SEQ ID NO:3:

ATGAAGGCCT TCGAGTCCCT CAAGTCCTTC

30

10

- (2) INFORMATION FOR SEQ ID NO:4:

- (i) SEQUENCE CHARACTERISTICS:
 (A) LENGTH: 30 base pairs
 (B) TYPE: nucleic acid
 (C) STRANDEDNESS: single
 (D) TOPOLOGY: linear

15

- (xi) SEQUENCE DESCRIPTION: SEQ ID NO:4:

AAACTCACGG TCGGTGCAGC GGCTCCTCAG

30

20

- (2) INFORMATION FOR SEQ ID NO:5:

- (i) SEQUENCE CHARACTERISTICS:
 (A) LENGTH: 10366 base pairs
 (B) TYPE: nucleic acid
 (C) STRANDEDNESS: single
 (D) TOPOLOGY: linear

25

- (ix) FEATURE:

- (A) NAME/KEY: CDS
 (B) LOCATION: 316..9748

30

- (xi) SEQUENCE DESCRIPTION: SEQ ID NO:5:

TTGCTGTGTG AGGCAGAACCTGCGGGGGCA GGGGCGGGCT GGTTCCTTGG CCAGCCATTG 60

35

GCAGAGTCCG CAGGCTAGGG CTGTCAATCA TGCTGGCCCG CGTGGCCCCG CCTCCGCCCG 120

CGCGGCCCCG CCTCCGCCCG CGCACGTCTG GGACGCAAGG CGCCGTGGGG GCTGCCGGGA 180

CGGGTCCAAG ATGGACGGCC GCTCAGGTTT TGCTTTTACC TGCGGCCAG AGCCCCATTC 240

40

ATTGCCCCCG TGCTGAGCGG CGCCGCGAGT CGGCCGAGG CCTCCGGGGA CTGCCGTGCC 300

GGGCGGGAGA CCGCC ATG GCG ACC CTG GAA AAG CTG ATG AAG GCC TTC GAG 351

Met Ala Thr Leu Glu Lys Leu Met Lys Ala Phe Glu
 1 5 10

45

TCC CTC AAG TCC TTC CAG CAG CAG CAG CAG CAG CAG CAG CAG CAG CAG 399

Ser Leu Lys Ser Phe Gln Gln Gln Gln Gln Gln Gln Gln Gln Gln
 15 20 25

CAG CAG CAG CAG CAG CAG CAG CAG CAG CAG CAA CAG CCG CCA CCG CCG 447

Gln Gln Gln Gln Gln Gln Gln Gln Gln Gln Gln Gln Gln Gln Gln
 30 35 40

50

CCG CCG CCG CCG CCG CCT CCT CAG CTT CCT CAG CCG CCG CCG CAG GCA 495

Pro Pro Pro Pro Pro Pro Pro Gln Leu Pro Gln Pro Pro Pro Pro Gln Ala
 45 50 55 60

CAG CCG CTG CTG CCT CAG CCG CAG CCG CCC CCG CCG CCG CCC CCG CCG 543

55

Gln Pro Leu Leu Pro Gln Pro Gln Pro Pro Pro Pro Pro Pro Pro Pro
 65 70 75

CCA CCC GGC CCG GCT GTG GCT GAG GAG CCG CTG CAC CGA CCA AAG AAA 591

EP 0 614 977 A2

	Pro	Pro	Gly	Pro	Ala	Val	Ala	Glu	Glu	Pro	Leu	His	Arg	Pro	Lys	Lys	
				80					85					90			
5	GAA	CTT	TCA	GCT	ACC	AAG	AAA	GAC	CGT	GTG	AAT	CAT	TGT	CTG	ACA	ATA	639
	Glu	Leu	Ser	Ala	Thr	Lys	Lys	Asp	Arg	Val	Asn	His	Cys	Leu	Thr	Ile	
			95					100					105				
	TGT	GAA	AAC	ATA	GTG	GCA	CAG	TCT	GTC	AGA	AAT	TCT	CCA	GAA	TTT	CAG	687
	Cys	Glu	Asn	Ile	Val	Ala	Gln	Ser	Val	Arg	Asn	Ser	Pro	Glu	Phe	Gln	
			110				115					120					
10	AAA	CTT	CTG	GGC	ATC	GCT	ATG	GAA	CTT	TTT	CTG	CTG	TGC	AGT	GAT	GAC	735
	Lys	Leu	Leu	Gly	Ile	Ala	Met	Glu	Leu	Phe	Leu	Leu	Cys	Ser	Asp	Asp	
						130					135					140	
	GCA	GAG	TCA	GAT	GTC	AGG	ATG	GTG	GCT	GAC	GAA	TGC	CTC	AAC	AAA	GTT	783
15	Ala	Glu	Ser	Asp	Val	Arg	Met	Val	Ala	Asp	Glu	Cys	Leu	Asn	Lys	Val	
					145					150					155		
	ATC	AAA	GCT	TTG	ATG	GAT	TCT	AAT	CTT	CCA	AGG	TTA	CAG	CTC	GAG	CTC	831
	Ile	Lys	Ala	Leu	Met	Asp	Ser	Asn	Leu	Pro	Arg	Leu	Gln	Leu	Glu	Leu	
				160				165					170				
20	TAT	AAG	GAA	ATT	AAA	AAG	AAT	GGT	GCC	CCT	CGG	AGT	TTG	CGT	GCT	GCC	879
	Tyr	Lys	Glu	Ile	Lys	Lys	Asn	Gly	Ala	Pro	Arg	Ser	Leu	Arg	Ala	Ala	
			175					180					185				
	CTG	TGG	AGG	TTT	GCT	GAG	CTG	GCT	CAC	CTG	GTT	CGG	CCT	CAG	AAA	TGC	927
	Leu	Trp	Arg	Phe	Ala	Glu	Leu	Ala	His	Leu	Val	Arg	Pro	Gln	Lys	Cys	
			190				195					200					
25	AGG	CCT	TAC	CTG	GTG	AAC	CTT	CTG	CCG	TGC	CTG	ACT	CGA	ACA	AGC	AAG	975
	Arg	Pro	Tyr	Leu	Val	Asn	Leu	Leu	Pro	Cys	Leu	Thr	Arg	Thr	Ser	Lys	
						210					215					220	
	AGA	CCC	GAA	GAA	TCA	GTC	CAG	GAG	ACC	TTG	GCT	GCA	GCT	GTT	CCC	AAA	1023
30	Arg	Pro	Glu	Glu	Ser	Val	Gln	Glu	Thr	Leu	Ala	Ala	Ala	Val	Pro	Lys	
					225					230					235		
	ATT	ATG	GCT	TCT	TTT	GGC	AAT	TTT	GCA	AAT	GAC	AAT	GAA	ATT	AAG	GTT	1071
	Ile	Met	Ala	Ser	Phe	Gly	Asn	Phe	Ala	Asn	Asp	Asn	Glu	Ile	Lys	Val	
				240				245					250				
35	TTG	TTA	AAG	GCC	TTC	ATA	GCG	AAC	CTG	AAG	TCA	AGC	TCC	CCC	ACC	ATT	1119
	Leu	Leu	Lys	Ala	Phe	Ile	Ala	Asn	Leu	Lys	Ser	Ser	Ser	Pro	Thr	Ile	
			255					260					265				
	CGG	CGG	ACA	GCG	GCT	GGA	TCA	GCA	GTG	AGC	ATC	TGC	CAG	CAC	TCA	AGA	1167
	Arg	Arg	Thr	Ala	Ala	Gly	Ser	Ala	Val	Ser	Ile	Cys	Gln	His	Ser	Arg	
			270				275					280					
40	AGG	ACA	CAA	TAT	TTC	TAT	AGT	TGG	CTA	CTA	AAT	GTG	CTC	TTA	GGC	TTA	1215
	Arg	Thr	Gln	Tyr	Phe	Tyr	Ser	Trp	Leu	Leu	Asn	Val	Leu	Leu	Gly	Leu	
						290					295				300		
	CTC	GTT	CCT	GTC	GAG	GAT	GAA	CAC	TCC	ACT	CTG	CTG	ATT	CTT	GGC	GTG	1263
45	Leu	Val	Pro	Val	Glu	Asp	Glu	His	Ser	Thr	Leu	Leu	Ile	Leu	Gly	Val	
					305					310					315		
	CTG	CTC	ACC	CTG	AGG	TAT	TTG	GTG	CCC	TTG	CTG	CAG	CAG	CAG	GTC	AAG	1311
	Leu	Leu	Thr	Leu	Arg	Tyr	Leu	Val	Pro	Leu	Leu	Gln	Gln	Gln	Val	Lys	
				320					325					330			
50	GAC	ACA	AGC	CTG	AAA	GGC	AGC	TTC	GGA	GTG	ACA	AGG	AAA	GAA	ATG	GAA	1359
	Asp	Thr	Ser	Leu	Lys	Gly	Ser	Phe	Gly	Val	Thr	Arg	Lys	Glu	Met	Glu	
			335					340					345				
	GTC	TCT	CCT	TCT	GCA	GAG	CAG	CTT	GTC	CAG	GTT	TAT	GAA	CTG	ACG	TTA	1407
	Val	Ser	Pro	Ser	Ala	Glu	Gln	Leu	Val	Gln	Val	Tyr	Glu	Leu	Thr	Leu	
			350				355					360					
55	CAT	CAT	ACA	CAG	CAC	CAA	GAC	CAC	AAT	GTT	GTG	ACC	GGA	GCC	CTG	GAG	1455
	His	His	Thr	Gln	His	Gln	Asp	His	Asn	Val	Val	Thr	Gly	Ala	Leu	Glu	

EP 0 614 977 A2

	365		370		375		380	
	CTG TTG CAG CAG CTC TTC AGA ACG CCT CCA CCC GAG CTT CTG CAA ACC							1503
5	Leu Leu Gln Gln Leu Phe Arg Thr Pro Pro Pro Glu Leu Leu Gln Thr		385		390		395	
	CTG ACC GCA GTC GGG GGC ATT GGG CAG CTC ACC GCT GCT AAG GAG GAG							1551
	Leu Thr Ala Val Gly Gly Ile Gly Gln Leu Thr Ala Ala Lys Glu Glu		400		405		410	
10	TCT GGT GGC CGA AGC CGT AGT GGG AGT ATT GTG GAA CTT ATA GCT GGA							1599
	Ser Gly Gly Arg Ser Arg Ser Gly Ser Ile Val Glu Leu Ile Ala Gly		415		420		425	
	GGG GGT TCC TCA TGC AGC CCT GTC CTT TCA AGA AAA CAA AAA GGC AAA							1647
15	Gly Gly Ser Ser Cys Ser Pro Val Leu Ser Arg Lys Gln Lys Gly Lys		430		435		440	
	GTG CTC TTA GGA GAA GAA GAA GCC TTG GAG GAT GAC TCT GAA TCG AGA							1695
	Val Leu Leu Gly Glu Glu Glu Ala Leu Glu Asp Asp Ser Glu Ser Arg		445		450		455	460
20	TCG GAT GTC AGC AGC TCT GCC TTA ACA GCC TCA GTG AAG GAT GAG ATC							1743
	Ser Asp Val Ser Ser Ser Ala Leu Thr Ala Ser Val Lys Asp Glu Ile		465		470		475	
	AGT GGA GAG CTG GCT GCT TCT TCA GGG GTT TCC ACT CCA GGG TCA GCA							1791
25	Ser Gly Glu Leu Ala Ala Ser Ser Gly Val Ser Thr Pro Gly Ser Ala		480		485		490	
	GGT CAT GAC ATC ATC ACA GAA CAG CCA CGG TCA CAG CAC ACA CTG CAG							1839
	Gly His Asp Ile Ile Thr Glu Gln Pro Arg Ser Gln His Thr Leu Gln		495		500		505	
30	GCG GAC TCA CTG GAT CTG GCC AGC TGT GAC TTG ACA AGC TCT GCC ACT							1887
	Ala Asp Ser Leu Asp Leu Ala Ser Cys Asp Leu Thr Ser Ser Ala Thr		510		515		520	
	GAT GGG GAT GAG GAG GAT ATC TTG AGC CAC AGC TCC AGC CAG GTC AGC							1935
	Asp Gly Asp Glu Glu Asp Ile Leu Ser His Ser Ser Ser Gln Val Ser		525		530		535	540
35	GCC GTC CCA TCT GAC CCT GCC ATG GAC CTG AAT GAT GGG ACC CAG GCC							1983
	Ala Val Pro Ser Asp Pro Ala Met Asp Leu Asn Asp Gly Thr Gln Ala		545		550		555	
40	TCG TCG CCC ATC AGC GAC AGC TCC CAG ACC ACC ACC GAA GGG CCT GAT							2031
	Ser Ser Pro Ile Ser Asp Ser Ser Gln Thr Thr Thr Glu Gly Pro Asp		560		565		570	
	TCA GCT GTT ACC CCT TCA GAC AGT TCT GAA ATT GTG TTA GAC GGT ACC							2079
	Ser Ala Val Thr Pro Ser Asp Ser Ser Glu Ile Val Leu Asp Gly Thr		575		580		585	
45	GAC AAC CAG TAT TTG GGC CTG CAG ATT GGA CAG CCC CAG GAT GAA GAT							2127
	Asp Asn Gln Tyr Leu Gly Leu Gln Ile Gly Gln Pro Gln Asp Glu Asp		590		595		600	
	GAG GAA GCC ACA CGT ATT CTT CCT CAT GAA CCC TCC CAG CCC TTC ACG							2175
	Glu Glu Ala Thr Gly Ile Leu Pro Asp Glu Ala Ser Glu Ala Phe Arg		605		610		615	620
50	AAC TCT TCC ATG GCC CTT CAA CAG GCA CAT TTA TTG AAA AAC ATG AGT							2223
	Asn Ser Ser Met Ala Leu Gln Gln Ala His Leu Leu Lys Asn Met Ser		625		630		635	
55	CAC TGC AGG CAG CCT TCT GAC AGC AGT GTT GAT AAA TTT GTG TTG AGA							2271
	His Cys Arg Gln Pro Ser Asp Ser Ser Val Asp Lys Phe Val Leu Arg		640		645		650	
	GAT GAA GCT ACT GAA CCG GGT GAT CAA GAA AAC AAG CCT TGC CGC ATC							2319
	Asp Glu Ala Thr Glu Pro Gly Asp Gln Glu Asn Lys Pro Cys Arg Ile							

EP 0 614 977 A2

		655				660						665					
		AAA GGT GAC ATT GGA CAG TCC ACT GAT GAT GAC TCT GCA CCT CTT GTC															2367
5		Lys Gly Asp Ile Gly Gln Ser Thr Asp Asp Asp Ser Ala Pro Leu Val															
		670				675						680					
		CAT TCT GTC CGC CTT TTA TCT GCT TCG TTT TTG CTA ACA GGG GGA AAA															2415
		His Ser Val Arg Leu Leu Ser Ala Ser Phe Leu Leu Thr Gly Gly Lys															
		685				690						695					700
10		AAT GTG CTG GTT CCG GAC AGG GAT GTG AGG GTC AGC GTG AAG GCC CTG															2463
		Asn Val Leu Val Pro Asp Arg Asp Val Arg Val Ser Val Lys Ala Leu															
						705						710					715
		GCC CTC AGC TGT GTG GGA GCA GCT GTG GCC CTC CAC CCG GAA TCT TTC															2511
15		Ala Leu Ser Cys Val Gly Ala Ala Val Ala Leu His Pro Glu Ser Phe															
						720						725					730
		TTC AGC AAA CTC TAT AAA GTT CCT CTT GAC ACC ACG GAA TAC CCT GAG															2559
		Phe Ser Lys Leu Tyr Lys Val Pro Leu Asp Thr Thr Glu Tyr Pro Glu															
						735						740					745
20		GAA CAG TAT GTC TCA GAC ATC TTG AAC TAC ATC GAT CAT GGA GAC CCA															2607
		Glu Gln Tyr Val Ser Asp Ile Leu Asn Tyr Ile Asp His Gly Asp Pro															
						750						755					760
		CAG GTT CGA GGA GCC ACT GCC ATT CTC TGT GGG ACC CTC ATC TGC TCC															2655
25		Gln Val Arg Gly Ala Thr Ala Ile Leu Cys Gly Thr Leu Ile Cys Ser															
						765						770					775
		ATC CTC AGC AGG TCC CGC TTC CAC GTG GGA GAT TGG ATG GGC ACC ATT															2703
		Ile Leu Ser Arg Ser Arg Phe His Val Gly Asp Trp Met Gly Thr Ile															
						785						790					795
30		AGA ACC CTC ACA GGA AAT ACA TTT TCT TTG GCG GAT TGC ATT CCT TTG															2751
		Arg Thr Leu Thr Gly Asn Thr Phe Ser Leu Ala Asp Cys Ile Pro Leu															
						800						805					810
		CTG CGG AAA ACA CTG AAG GAT GAG TCT TCT GTT ACT TGC AAG TTA GCT															2799
		Leu Arg Lys Thr Leu Lys Asp Glu Ser Ser Val Thr Cys Lys Leu Ala															
						815						820					825
35		TGT ACA GCT GTG AGG AAC TGT GTC ATG AGT CTC TGC AGC AGC AGC TAC															2847
		Cys Thr Ala Val Arg Asn Cys Val Met Ser Leu Cys Ser Ser Ser Tyr															
						830						835					840
		AGT GAG TTA GGA CTG CAG CTG ATC ATC GAT GTG CTG ACT CTG AGG AAC															2895
40		Ser Glu Leu Gly Leu Gln Leu Ile Ile Asp Val Leu Thr Leu Arg Asn															
						845						850					855
		AGT TCC TAT TGG CTG GTG AGG ACA GAG CTT CTG GAA ACC CTT GCA GAG															2943
		Ser Ser Tyr Trp Leu Val Arg Thr Glu Leu Leu Glu Thr Leu Ala Glu															
						865						870					875
45		ATT GAC TTC AGG CTG GTG AGC TTT TTG GAG GCA AAA GCA GAA AAC TTA															2991
		Ile Asp Phe Arg Leu Val Ser Phe Leu Glu Ala Lys Ala Glu Asn Leu															
						880						885					890
		CAC AGA GGG GCT CAT CAT TAT ACA GGG CTT TTA AAA CTG CAA GAA CGA															3039
		His Arg Gly Ala His His Tyr Thr Gly Leu Leu Lys Leu Gln Glu Arg															
						895						900					905
50		GTG CTC AAT AAT GTT GTC ATC CAT TTG CTT GGA GAT GAA GAC CCC AGG															3087
		Val Leu Asn Asn Val Val Ile His Leu Leu Gly Asp Glu Asp Pro Arg															
						910						915					920
		GTG CGA CAT GTT GCC GCA GCA TCA CTA ATT AGG CTT GTC CCA AAG CTG															3135
55		Val Arg His Val Ala Ala Ala Ser Leu Ile Arg Leu Val Pro Lys Leu															
						925						930					935
		TTT TAT AAA TGT GAC CAA GGA CAA GCT GAT CCA GTA GTG GCC GTG GCA															3183
		Phe Tyr Lys Cys Asp Gln Gly Gln Ala Asp Pro Val Val Ala Val Ala															

EP 0 614 977 A2

				945				950					955				
5				AGA GAT CAA AGC AGT GTT TAC CTG AAA CTT CTC ATG CAT GAG ACG CAG													3231
				Arg Asp Gln Ser Ser Val Tyr Leu Lys Leu Leu Met His Glu Thr Gln													
				960				965					970				
				CCT CCA TCT CAT TTC TCC GTC AGC ACA ATA ACC AGA ATA TAT AGA GGC													3279
				Pro Pro Ser His Phe Ser Val Ser Thr Ile Thr Arg Ile Tyr Arg Gly													
				975				980					985				
10				TAT AAC CTA CTA CCA AGC ATA ACA GAC GTC ACT ATG GAA AAT AAC CTT													3327
				Tyr Asn Leu Leu Pro Ser Ile Thr Asp Val Thr Met Glu Asn Asn Leu													
				990				995					1000				
				TCA AGA GTT ATT GCA GCA GTT TCT CAT GAA CTA ATC ACA TCA ACC ACC													3375
				Ser Arg Val Ile Ala Ala Val Ser His Glu Leu Ile Thr Ser Thr Thr													
15				1005				1010				1015					1020
				AGA GCA CTC ACA TTT GGA TGC TGT GAA GCT TTG TGT CTT CTT TCC ACT													3423
				Arg Ala Leu Thr Phe Gly Cys Cys Glu Ala Leu Cys Leu Leu Ser Thr													
				1025				1030									1035
20				GCC TTC CCA GTT TGC ATT TGG AGT TTA GGT TGG CAC TGT GGA GTG CCT													3471
				Ala Phe Pro Val Cys Ile Trp Ser Leu Gly Trp His Cys Gly Val Pro													
				1040				1045									1050
				CCA CTG AGT GCC TCA GAT GAG TCT AGG AAG AGC TGT ACC GTT GGG ATG													3519
				Pro Leu Ser Ala Ser Asp Glu Ser Arg Lys Ser Cys Thr Val Gly Met													
				1055				1060									1065
25				GCC ACA ATG ATT CTG ACC CTG CTC TCG TCA GCT TGG TTC CCA TTG GAT													3567
				Ala Thr Met Ile Leu Thr Leu Leu Ser Ser Ala Trp Phe Pro Leu Asp													
				1070				1075									1080
				CTC TCA GCC CAT CAA GAT GCT TTG ATT TTG GCC GGA AAC TTG CTT GCA													3615
				Leu Ser Ala His Gln Asp Ala Leu Ile Leu Ala Gly Asn Leu Leu Ala													
30				1085				1090					1095				1100
				GCC AGT GCT CCC AAA TCT CTG AGA AGT TCA TGG GCC TCT GAA GAA GAA													3663
				Ala Ser Ala Pro Lys Ser Leu Arg Ser Ser Trp Ala Ser Glu Glu Glu													
				1105				1110									1115
35				GCC AAC CCA GCA GCC ACC AAG CAA GAG GAG GTC TGG CCA GCC CTG GGG													3711
				Ala Asn Pro Ala Ala Thr Lys Gln Glu Glu Val Trp Pro Ala Leu Gly													
				1120				1125									1130
				GAC CGG GCC CTG GTG CCC ATG GTG GAG CAG CTC TTC TCT CAC CTG CTG													3759
				Asp Arg Ala Leu Val Pro Met Val Glu Gln Leu Phe Ser His Leu Leu													
40				1135				1140									1145
				AAG GTG ATT AAC ATT TGT GCC CAC GTC CTG GAT GAC GTG GCT CCT GGA													3807
				Lys Val Ile Asn Ile Cys Ala His Val Leu Asp Asp Val Ala Pro Gly													
				1150				1155									1160
				CCC GCA ATA AAG GCA GCC TTG CCT TCT CTA ACA AAC CCC CCT TCT CTA													3855
				Pro Ala Ile Lys Ala Ala Leu Pro Ser Leu Thr Asn Pro Pro Ser Leu													
45				1165				1170					1175				1180
				AGT CCC ATC CGA CGA AAG GGG AAG GAG AAA GAA CCA GGA GAA CAA GCA													3903
				Ser Pro Ile Arg Arg Lys Gly Lys Glu Lys Glu Pro Gly Glu Gln Ala													
				1185				1190									1195
50				TCT GTA CCG TTG AGT CCC AAG AAA GGC AGT GAG GCC AGT GCA GCT TCT													3951
				Ser Val Pro Leu Ser Pro Lys Lys Gly Ser Glu Ala Ser Ala Ala Ser													
				1200				1205									1210
				AGA CAA TCT GAT ACC TCA GGT CCT GTT ACA ACA AGT AAA TCC TCA TCA													3999
				Arg Gln Ser Asp Thr Ser Gly Pro Val Thr Thr Ser Lys Ser Ser Ser													
55				1215				1220									1225
				CTG GGG AGT TTC TAT CAT CTT CCT TCA TAC CTC AGA CTG CAT GAT GTC													4047
				Leu Gly Ser Phe Tyr His Leu Pro Ser Tyr Leu Arg Leu His Asp Val													

EP 0 614 977 A2

	1230	1235	1240	
	CTG AAA GCT ACA CAC GCT AAC TAC AAG GTC ACG CTG GAT CTT CAG AAC			4095
	Leu Lys Ala Thr His Ala Asn Tyr Lys Val Thr Leu Asp Leu Gln Asn			
5	1245	1250	1255	1260
	AGC ACG GAA AAG TTT GGA GGG TTT CTC CGC TCA GCC TTG GAT GTT CTT			4143
	Ser Thr Glu Lys Phe Gly Gly Phe Leu Arg Ser Ala Leu Asp Val Leu			
		1265	1270	1275
10	TCT CAG ATA CTA GAG CTG GCC ACA CTG CAG GAC ATT GGG AAG TGT GTT			4191
	Ser Gln Ile Leu Glu Leu Ala Thr Leu Gln Asp Ile Gly Lys Cys Val			
		1280	1285	1290
	GAA GAG ATC CTA GGA TAC CTG AAA TCC TGC TTT AGT CGA GAA CCA ATG			4239
	Glu Glu Ile Leu Gly Tyr Leu Lys Ser Cys Phe Ser Arg Glu Pro Met			
15		1295	1300	1305
	ATG GCA ACT GTT TGT GTT CAA CAA TTG TTG AAG ACT CTC TTT GGC ACA			4287
	Met Ala Thr Val Cys Val Gln Gln Leu Leu Lys Thr Leu Phe Gly Thr			
		1310	1315	1320
20	AAC TTG GCC TCC CAG TTT GAT GGC TTA TCT TCC AAC CCC AGC AAG TCA			4335
	Asn Leu Ala Ser Gln Phe Asp Gly Leu Ser Ser Asn Pro Ser Lys Ser			
		1325	1330	1335
	CAA GGC CGA GCA CAG CGC CTT GGC TCC TCC AGT GTG AGG CCA GGC TTG			4383
	Gln Gly Arg Ala Gln Arg Leu Gly Ser Ser Ser Val Arg Pro Gly Leu			
		1345	1350	1355
25	TAC CAC TAC TGC TTC ATG GCC CCG TAC ACC CAC TTC ACC CAG GCC CTC			4431
	Tyr His Tyr Cys Phe Met Ala Pro Tyr Thr His Phe Thr Gln Ala Leu			
		1360	1365	1370
	GCT GAC GCC AGC CTG AGG AAC ATG GTG CAG GCG GAG CAG GAG AAC GAC			4479
	Ala Asp Ala Ser Leu Arg Asn Met Val Gln Ala Glu Gln Glu Asn Asp			
30		1375	1380	1385
	ACC TCG GGA TGG TTT GAT GTC CTC CAG AAA GTG TCT ACC CAG TTG AAG			4527
	Thr Ser Gly Trp Phe Asp Val Leu Gln Lys Val Ser Thr Gln Leu Lys			
		1390	1395	1400
35	ACA AAC CTC ACG AGT GTC ACA AAG AAC CGT GCA GAT AAG AAT GCT ATT			4575
	Thr Asn Leu Thr Ser Val Thr Lys Asn Arg Ala Asp Lys Asn Ala Ile			
		1405	1410	1415
	CAT AAT CAC ATT CGT TTG TTT GAA CCT CTT GTT ATA AAA GCT TTA AAA			4623
	His Asn His Ile Arg Leu Phe Glu Pro Leu Val Ile Lys Ala Lys Lys			
40		1425	1430	1435
	CAG TAC ACG ACT ACA ACA TGT GTG CAG TTA CAG AAG CAG GTT TTA GAT			4671
	Gln Tyr Thr Thr Thr Cys Val Gln Leu Gln Lys Gln Val Leu Asp			
		1440	1445	1450
45	TTG CTG GCG CAG CTG GTT CAG TTA CGG GTT AAT TAC TGT CTT CTG GAT			4719
	Leu Leu Ala Gln Leu Val Gln Leu Arg Val Asn Tyr Cys Leu Leu Asp			
		1455	1460	1465
	TCA GAT CAG GTG TTT ATT GGC TTT GTA TTG AAA CAG TTT GAA TAC ATT			4767
	Ser Asp Gln Val Phe Ile Gly Phe Val Leu Lys Gln Phe Glu Tyr Ile			
		1470	1475	1480
50	GAA GTG GGC CAG TTC AGG GAA TCA GAG GCA ATC ATT CCA AAC ATC TTT			4815
	Glu Val Gly Gln Phe Arg Glu Ser Glu Ala Ile Ile Pro Asn Ile Phe			
		1485	1490	1500
	TTC TTC TTG GTA TTA CTA TCT TAT GAA CGC TAT CAT TCA AAA CAG ATC			4863
	Phe Phe Leu Val Leu Leu Ser Tyr Glu Arg Tyr His Ser Lys Gln Ile			
55		1505	1510	1515
	ATT GGA ATT CCT AAA ATC ATT CAG CTC TGT GAT GGC ATC ATG GCC AGT			4911
	Ile Gly Ile Pro Lys Ile Ile Gln Leu Cys Asp Gly Ile Met Ala Ser			

EP 0 614 977 A2

	1520	1525	1530	
5	GGA AGG AAG GCT GTG ACA CAT GCC ATA CCG GCT CTG CAG CCC ATA GTC Gly Arg Lys Ala Val Thr His Ala Ile Pro Ala Leu Gln Pro Ile Val 1535 1540 1545			4959
10	CAC GAC CTC TTT GTA TTA AGA GGA ACA AAT AAA GCT GAT GCA GGA AAA His Asp Leu Phe Val Leu Arg Gly Thr Asn Lys Ala Asp Ala Gly Lys 1550 1555 1560			5007
	GAG CTT GAA ACC CAA AAA GAG GTG GTG GTG TCA ATG TTA CTG AGA CTC Glu Leu Glu Thr Gln Lys Glu Val Val Val Ser Met Leu Leu Arg Leu 1565 1570 1575 1580			5055
15	ATC CAG TAC CAT CAG GTG TTG GAG ATG TTC ATT CTT GTC CTG CAG CAG Ile Gln Tyr His Gln Val Leu Glu Met Phe Ile Leu Val Leu Gln Gln 1585 1590 1595			5103
	TGC CAC AAG GAG AAT GAA GAC AAG TGG AAG CGA CTG TCT CGA CAG ATA Cys His Lys Glu Asn Glu Asp Lys Trp Lys Arg Leu Ser Arg Gln Ile 1600 1605 1610			5151
20	GCT GAC ATC ATC CTC CCA ATG TTA GCC AAA CAG CAG ATG CAC ATT GAC Ala Asp Ile Ile Leu Pro Met Leu Ala Lys Gln Gln Met His Ile Asp 1615 1620 1625			5199
25	TCT CAT GAA GCC CTT GGA GTG TTA AAT ACA TTA TTT GAG ATT TTG GCC Ser His Glu Ala Leu Gly Val Leu Asn Thr Leu Phe Glu Ile Leu Ala 1630 1635 1640			5247
	CCT TCC TCC CTC CGT CCG GTA GAC ATG CTT TTA CGG AGT ATG TTC GTC Pro Ser Ser Leu Arg Pro Val Asp Met Leu Leu Arg Ser Met Phe Val 1645 1650 1655 1660			5295
30	ACT CCA AAC ACA ATG GCG TCC GTG AGC ACT GTT CAA CTG TGG ATA TCG Thr Pro Asn Thr Met Ala Ser Val Ser Thr Val Gln Leu Trp Ile Ser 1665 1670 1675			5343
	GGA ATT CTG GCC ATT TTG AGG GTT CTG ATT TCC CAG TCA ACT GAA GAT Gly Ile Leu Ala Ile Leu Arg Val Leu Ile Ser Gln Ser Thr Glu Asp 1680 1685 1690			5391
35	ATT GTT CTT TCT CGT ATT CAG GAG CTC TCC TTC TCT CCG TAT TTA ATC Ile Val Leu Ser Arg Ile Gln Glu Leu Ser Phe Ser Pro Tyr Leu Ile 1695 1700 1705			5439
40	TCC TGT ACA GTA ATT AAT AGG TTA AGA GAT GGG GAC AGT ACT TCA ACG Ser Cys Thr Val Ile Asn Arg Leu Arg Asp Gly Asp Ser Thr Ser Thr 1710 1715 1720			5487
	CTA GAA GAA CAC AGT GAA GGG AAA CAA ATA AAG AAT TTG CCA GAA GAA Leu Glu Glu His Ser Glu Gly Lys Gln Ile Lys Asn Leu Pro Glu Glu 1725 1730 1735 1740			5535
45	ACA TTT TCA AGG TTT CTA TTA CAA CTG GTT GGT ATT CTT TTA GAA GAC Thr Phe Ser Arg Phe Leu Leu Gln Leu Val Gly Ile Leu Leu Glu Asp 1745 1750 1755			5583
	ATT GTT ACA AAA CAG CTG AAG GTG GAA ATG AGT GAG CAG CAA CAT ACT Ile Val Thr Lys Gln Leu Lys Val Glu Met Ser Glu Gln Gln His Thr 1760 1765 1770			5631
50	TTC TAT TGC CAG GAA CTA GGC ACA CTG CTA ATG TGT CTG ATC CAC ATC Phe Tyr Cys Glr Glu Leu Gly Thr Leu Leu Met Cys Leu Ile His Ile 1775 1780 1785			5679
55	TTC AAG TCT GGA ATG TTC CCG AGA ATC ACA GCA GCT GCC ACT AGG CTG Phe Lys Ser Gly Met Phe Arg Arg Ile Thr Ala Ala Thr Arg Leu 1790 1795 1800			5727
	TTC CGC AGT GAT GGC TGT GGC GGC AGT TTC TAC ACC CTG GAC AGC TTG Phe Arg Ser Asp Gly Cys Gly Gly Ser Phe Tyr Thr Leu Asp Ser Leu 1805 1810 1815			5775

EP 0 614 977 A2

	1805		1810		1815		1820	
	AAC TTG CGG GCT CGT TCC ATG ATC ACC ACC CAC CCG GCC CTG GTG CTG							5823
5	Asn Leu Arg Ala Arg Ser Met Ile Thr Thr His Pro Ala Leu Val Leu		1825		1830		1835	
	CTC TGG TGT CAG ATA CTG CTG CTT GTC AAC CAC ACC GAC TAC CGC TGG							5871
	Leu Trp Cys Gln Ile Leu Leu Leu Val Asn His Thr Asp Tyr Arg Trp		1840		1845		1850	
10	TGG GCA GAA GTG CAG CAG ACC CCG AAA AGA CAC AGT CTG TCC AGC ACA							5919
	Trp Ala Glu Val Gln Gln Thr Pro Lys Arg His Ser Leu Ser Ser Thr		1855		1860		1865	
	AAG TTA CTT AGT CCC CAG ATG TCT GGA GAA GAG GAG GAT TCT GAC TTG							5967
15	Lys Leu Leu Ser Pro Gln Met Ser Gly Glu Glu Glu Asp Ser Asp Leu		1870		1875		1880	
	GCA GCC AAA CTT GGA ATG TGC AAT AGA GAA ATA GTA CGA AGA GGG GCT							6015
	Ala Ala Lys Leu Gly Met Cys Asn Arg Glu Ile Val Arg Arg Gly Ala		1885		1890		1895	1900
20	CTC ATT CTC TTC TGT GAT TAT GTC TGT CAG AAC CTC CAT GAC TCC GAG							6063
	Leu Ile Leu Phe Cys Asp Tyr Val Cys Gln Asn Leu His Asp Ser Glu		1905		1910		1915	
	CAC TTA ACG TGG CTC ATT GTA AAT CAC ATT CAA GAT CTG ATC AGC CTT							6111
25	His Leu Thr Trp Leu Ile Val Asn His Ile Gln Asp Leu Ile Ser Leu		1920		1925		1930	
	TCC CAC GAG CCT CCA GTA CAG GAC TTC ATC AGT GCC GTT CAT CGG AAC							6159
	Ser His Glu Pro Pro Val Gln Asp Phe Ile Ser Ala Val His Arg Asn		1935		1940		1945	
30	TCT GCT GCC AGC GGC CTG TTC ATC CAG GCA ATT CAG TCT CGT TGT GAA							6207
	Ser Ala Ala Ser Gly Leu Phe Ile Gln Ala Ile Gln Ser Arg Cys Glu		1950		1955		1960	
	AAC CTT TCA ACT CCA ACC ATG CTG AAG AAA ACT CTT CAG TGC TTG GAG							6255
	Asn Leu Ser Thr Pro Thr Met Leu Lys Lys Thr Leu Gln Cys Leu Glu		1965		1970		1975	1980
35	GGG ATC CAT CTC AGC CAG TCG GGA GCT GTG CTC ACG CTG TAT GTG GAC							6303
	Gly Ile His Leu Ser Gln Ser Gly Ala Val Leu Thr Leu Tyr Val Asp		1985		1990		1995	
	AGG CTT CTG TGC ACC CCT TTC CGT GTG CTG GCT CGC ATG GTC GAC ATC							6351
40	Arg Leu Leu Cys Thr Pro Phe Arg Val Leu Ala Arg Met Val Asp Ile		2000		2005		2010	
	CTT GCT TGT CGC CGG GTA GAA ATG CTT CTG GCT GCA AAT TTA CAG AGC							6399
	Leu Ala Cys Arg Arg Val Glu Met Leu Leu Ala Ala Asn Leu Gln Ser		2015		2020		2025	
45	AGC ATG GCC CAG TTG CCA ATG GAA GAA CTC AAC AGA ATC CAG GAA TAC							6447
	Ser Met Ala Gln Leu Pro Met Glu Glu Leu Asn Arg Ile Gln Glu Tyr		2030		2035		2040	
	CTT CAG AGC AGC GGG CTC GCT CAG AGA CAC CAA AGG CTC TAT TCC CTG							6495
50	Leu Gln Ser Ser Gly Leu Ala Gln Arg His Gln Arg Leu Tyr Ser Leu		2045		2050		2055	2060
	CTG GAC AGG TTT CGT CTC TCC ACC ATG CAA GAC TCA CTT AGT CCC TCT							6543
	Leu Asp Arg Phe Arg Leu Ser Thr Met Gln Asp Ser Leu Ser Pro Ser		2065		2070		2075	
	CCT CCA GTC TCT TCC CAC CCG CTG GAC GGG GAT GGG CAC GTG TCA CTG							6591
55	Pro Pro Val Ser Ser His Pro Leu Asp Gly Asp Gly His Val Ser Leu		2080		2085		2090	
	GAA ACA GTG AGT CCG GAC AAA GAC TGG TAC GTT CAT CTT GTC AAA TCC							6639
	Glu Thr Val Ser Pro Asp Lys Asp Trp Tyr Val His Leu Val Lys Ser							

EP 0 614 977 A2

	2095	2100	2105	
5	CAG TGT TGG ACC AGG TCA GAT TCT GCA CTG CTG GAA GGT GCA GAG CTG Gln Cys Trp Thr Arg Ser Asp Ser Ala Leu Leu Glu Gly Ala Glu Leu 2110 2115 2120			6687
10	GTG AAT CGG ATT CCT GCT GAA GAT ATG AAT GCC TTC ATG ATG AAC TCG Val Asn Arg Ile Pro Ala Glu Asp Met Asn Ala Phe Met Met Asn Ser 2125 2130 2135 2140			6735
15	GAG TTC AAC CTA AGC CTG CTA GCT CCA TGC TTA AGC CTA GGG ATG AGT Glu Phe Asn Leu Ser Leu Leu Ala Pro Cys Leu Ser Leu Gly Met Ser 2145 2150 2155			6783
20	GAA ATT TCT GGT GGC CAG AAG AGT GCC CTT TTT GAA GCA GCC CGT GAG Glu Ile Ser Gly Gly Gln Lys Ser Ala Leu Phe Glu Ala Ala Arg Glu 2160 2165 2170			6831
25	GTG ACT CTG GCC CGT GTG AGC GGC ACC GTG CAG CAG CTC CCT GCT GTC Val Thr Leu Ala Arg Val Ser Gly Thr Val Gln Gln Leu Pro Ala Val 2175 2180 2185			6879
30	CAT CAT GTC TTC CAG CCC GAG CTG CCT GCA GAG CCG GCG GCC TAC TGG His His Val Phe Gln Pro Glu Leu Pro Ala Glu Pro Ala Ala Tyr Trp 2190 2195 2200			6927
35	AGC AAG TTG AAT GAT CTG TTT GGG GAT GCT GCA CTG TAT CAG TCC CTG Ser Lys Leu Asn Asp Leu Phe Gly Asp Ala Ala Leu Tyr Gln Ser Leu 2205 2210 2215 2220			6975
40	CCC ACT CTG GCC CGG GCC CTG GCA CAG TAC CTG GTG GTG GTC TCC AAA Pro Thr Leu Ala Arg Ala Leu Ala Gln Tyr Leu Val Val Val Ser Lys 2225 2230 2235			7023
45	CTG CCC AGT CAT TTG CAC CTT CCT CCT GAG AAA GAG AAG GAC ATT GTG Leu Pro Ser His Leu His Leu Pro Pro Glu Lys Glu Lys Asp Ile Val 2240 2245 2250			7071
50	AAA TTC GTG GTG GCA ACC CTT GAG GCC CTG TCC TGG CAT TTG ATC CAT Lys Phe Val Val Ala Thr Leu Glu Ala Leu Ser Trp His Leu Ile His 2255 2260 2265			7119
55	GAG CAG ATC CCG CTG AGT CTG GAT CTC CAG GCA GGG CTG GAC TGC TGC Glu Gln Ile Pro Leu Ser Leu Asp Leu Gln Ala Gly Leu Asp Cys Cys 2270 2275 2280			7167
60	TGC CTG GCC CTG CAG CTG CCT GGC CTC TGG AGC GTG GTC TCC TCC ACA Cys Leu Ala Leu Gln Leu Pro Gly Leu Trp Ser Val Val Ser Ser Thr 2285 2290 2295 2300			7215
65	GAG TTT GTG ACC CAC GCC TGC TCC CTC ATC TAC TGT GTG CAC TTC ATC Glu Phe Val Thr His Ala Cys Ser Leu Ile Tyr Cys Val His Phe Ile 2305 2310 2315			7263
70	CTG GAG GCC GTT GCA GTG CAG CCT GGA GAG CAG CTT CTT AGT CCA GAA Leu Glu Ala Val Ala Val Gln Pro Gly Glu Gln Leu Leu Ser Pro Glu 2320 2325 2330			7311
75	AGA AGG ACA AAT ACC CCA AAA GCC ATC AGC GAG GAG GAG GAG GAA GTA Arg Arg Thr Asn Thr Pro Lys Ala Ile Ser Glu Glu Glu Glu Val 2335 2340 2345			7359
80	GAT CCA AAC ACA CAG AAT CCT AAG TAT ATC ACT GCA GCC TGT GAG ATG Asp Pro Asn Thr Gln Asn Pro Lys Tyr Ile Thr Ala Ala Cys Glu Met 2350 2355 2360			7407
85	GTG GCA GAA ATG GTG GAG TCT CTG CAG TCG GTG TTG GCC TTG GGT CAT Val Ala Glu Met Val Glu Ser Leu Gln Ser Val Leu Ala Leu Gly His 2365 2370 2375 2380			7455
90	AAA AGG AAT AGC GGC GTG CCG GCG TTT CTC ACG CCA TTG CTC AGG AAC Lys Arg Asn Ser Gly Val Pro Ala Phe Leu Thr Pro Leu Leu Arg Asn			7503

EP 0 614 977 A2

	2385	2390	2395	
5	ATC ATC ATC AGC CTG GCC CGC CTG CCC CTT GTC AAC AGC TAC ACA CGT Ile Ile Ile Ser Leu Ala Arg Leu Pro Leu Val Asn Ser Tyr Thr Arg 2400 2405 2410			7551
10	GTG CCC CCA CTG GTG TGG AAG CTT GGA TGG TCA CCC AAA CCG GGA GGG Val Pro Pro Leu Val Trp Lys Leu Gly Trp Ser Pro Lys Pro Gly Gly 2415 2420 2425			7599
15	GAT TTT GGC ACA GCA TTC CCT GAG ATC CCC GTG GAG TTC CTC CAG GAA Asp Phe Gly Thr Ala Phe Pro Glu Ile Pro Val Glu Phe Leu Gln Glu 2430 2435 2440			7647
20	AAG GAA GTC TTT AAG GAG TTC ATC TAC CGC ATC AAC ACA CTA GGC TGG Lys Glu Val Phe Lys Glu Phe Ile Tyr Arg Ile Asn Thr Leu Gly Trp 2445 2450 2455 2460			7695
25	ACC AGT CGT ACT CAG TTT GAA GAA ACT TGG GCC ACC CTC CTT GGT GTC Thr Ser Arg Thr Gln Phe Glu Glu Thr Trp Ala Thr Leu Leu Gly Val 2465 2470 2475			7743
30	CTG GTG ACC CAG CCC CTC GTG ATG GAG CAG GAG GAG AGC CCA CCA GAA Leu Val Thr Gln Pro Leu Val Met Glu Gln Glu Glu Ser Pro Pro Glu 2480 2485 2490			7791
35	GAA GAC ACA GAG AGG ACC CAG ATC AAC GTC CTG GCC GTG CAG GCC ATC Glu Asp Thr Glu Arg Thr Gln Ile Asn Val Leu Ala Val Gln Ala Ile 2495 2500 2505			7839
40	ACC TCA CTG GTG CTC ACT GCA ATG ACT GTG CCT GTG GCC GGC AAC CCA Thr Ser Leu Val Leu Ser Ala Met Thr Val Pro Val Ala Gly Asn Pro 2510 2515 2520			7887
45	GCT GTA AGC TGC TTG GAG CAG CAG CCC CGG AAC AAG CCT CTG AAA GCT Ala Val Ser Cys Leu Glu Gln Gln Pro Arg Asn Lys Pro Leu Lys Ala 2525 2530 2535 2540			7935
50	CTC GAC ACC AGG TTT GGG AGG AAG CTG AGC ATT ATC AGA GGG ATT GTG Leu Asp Thr Arg Phe Gly Arg Lys Leu Ser Ile Ile Arg Gly Ile Val 2545 2550 2555			7983
55	GAG CAA GAG ATT CAA GCA ATG GTT TCA AAG AGA GAG AAT ATT GCC ACC Glu Gln Glu Ile Gln Ala Met Val Ser Lys Arg Glu Asn Ile Ala Thr 2560 2565 2570			8031
60	CAT CAT TTA TAT CAG GCA TGG GAT CCT GTC CCT TCT CTG TCT CCG GCT His His Leu Tyr Gln Ala Trp Asp Pro Val Pro Ser Leu Ser Pro Ala 2575 2580 2585			8079
65	ACT ACA GGT GCC CTC ATC AGC CAC GAG AAG CTG CTG CTA CAG ATC AAC Thr Thr Gly Ala Leu Ile Ser His Glu Lys Leu Leu Leu Gln Ile Asn 2590 2595 2600			8127
70	CCC GAG CGG GAG CTG GGG AGC ATG AGC TAC AAA CTC GGC CAG GTG TCC Pro Glu Arg Glu Leu Gly Ser Met Ser Tyr Lys Leu Gly Gln Val Ser 2605 2610 2615 2620			8175
75	ATA CAC TCC CTG TGG CTG GGG AAC AGC ATC ACA CCC CTG AGG GAG GAG Ile His Ser Val Trp Leu Gly Asn Ser Ile Thr Pro Leu Arg Glu Glu 2625 2630 2635			8223
80	GAA TGG GAC GAG GAA GAG GAG GAG GAG GCC GAC GCC CCT GCA CCT TCG Glu Trp Asp Glu Glu Glu Glu Glu Glu Ala Asp Ala Pro Ala Pro Ser 2640 2645 2650			8271
85	TCA CCA CCC ACG TCT CCA GTC AAC TCC AGG AAA CAC CGG GCT GGA GTT Ser Pro Thr Ser Pro Val Asn Ser Arg Lys His Arg Ala Gly Val 2655 2660 2665			8319
90	GAC ATC CAC TCC TGT TCG CAG TTT TTG CTT GAG TTG TAC AGC CGC TGG Asp Ile His Ser Cys Ser Gln Phe Leu Leu Glu Leu Tyr Ser Arg Trp			8367

EP 0 614 977 A2

	2670	2675	2680	
5	ATC CTG CCG TCC AGC TCA GCC AGG AGG ACC CCG GCC ATC CTG ATC AGT Ile Leu Pro Ser Ser Ser Ala Arg Arg Thr Pro Ala Ile Leu Ile Ser 2685 2690 2695 2700			8415
	GAG GTG GTC AGA TCC CTT CTA GTG GTC TCA GAC TTG TTC ACC GAG CGC Glu Val Val Arg Ser Leu Leu Val Val Ser Asp Leu Phe Thr Glu Arg 2705 2710 2715			8463
10	AAC CAG TTT GAG CTG ATG TAT GTG ACG CTG ACA GAA CTG CGA AGG GTG Asn Gln Phe Glu Leu Met Tyr Val Thr Leu Thr Glu Leu Arg Arg Val 2720 2725 2730			8511
15	CAC CCT TCA GAA GAC GAG ATC CTC GCT CAG TAC CTG GTG CCT GCC ACC His Pro Ser Glu Asp Glu Ile Leu Ala Gln Tyr Leu Val Pro Ala Thr 2735 2740 2745			8559
	TGC AAG GCA GCT GCC GTC CTT GGG ATG GAC AAG GCC GTG GCG GAG CCT Cys Lys Ala Ala Val Leu Gly Met Asp Lys Ala Val Ala Glu Pro 2750 2755 2760			8607
20	GTC AGC CGC CTG CTG GAG AGC ACG CTC AGG AGC AGC CAC CTG CCC AGC Val Ser Arg Leu Leu Glu Ser Thr Leu Arg Ser Ser His Leu Pro Ser 2765 2770 2775 2780			8655
	AGG GTT GGA GCC CTG CAC GGC ATC CTC TAT GTG CTG GAG TGC GAC CTG Arg Val Gly Ala Leu His Gly Ile Leu Tyr Val Leu Glu Cys Asp Leu 2785 2790 2795			8703
25	CTG GAC GAC ACT GCC AAG CAG CTC ATC CCG GTC ATC AGC GAC TAT CTC Leu Asp Asp Thr Ala Lys Gln Leu Ile Pro Val Ile Ser Asp Tyr Leu 2800 2805 2810			8751
30	CTC TCC AAC CTG AAA GGG ATC GCC CAC TGC GTG AAC ATT CAC AGC CAG Leu Ser Asn Leu Lys Gly Ile Ala His Cys Val Asn Ile His Ser Gln 2815 2820 2825			8799
	CAG CAC GTA CTG GTC ATG TGT GCC ACT GCG TTT TAC CTC ATT GAG AAC Gln His Val Leu Val Met Cys Ala Thr Ala Phe Tyr Leu Ile Glu Asn 2830 2835 2840			8847
35	TAT CCT CTG GAC GTA GGG CCG GAA TTT TCA GCA TCA ATA ATA CAG ATG Tyr Pro Leu Asp Val Gly Pro Glu Phe Ser Ala Ser Ile Ile Gln Met 2845 2850 2855 2860			8895
	TGT GGG GTG ATG CTG TCT GGA AGT GAG GAG TCC ACC CCC TCC ATC ATT Cys Gly Val Met Leu Ser Gly Ser Glu Glu Ser Thr Pro Ser Ile Ile 2865 2870 2875			8943
40	TAC CAC TGT GCC CTC AGA GGC CTG GAG CGC CTC CTG CTC TCT GAG CAG Tyr His Cys Ala Leu Arg Gly Leu Glu Arg Leu Leu Leu Ser Glu Gln 2880 2885 2890			8991
45	CTC TCC CGC CTG GAT GCA GAA TCG CTG GTC AAG CTG AGT GTG GAC AGA Leu Ser Arg Leu Asp Ala Glu Ser Leu Val Lys Leu Ser Val Asp Arg 2895 2900 2905			9039
	GTG AAC GTG CAC AGC CCG CAC CGG GCC ATG GCG GCT CTG GGC CTG ATG Val Asn Val His Ser Pro His Arg Ala Met Ala Ala Leu Gly Leu Met 2910 2915 2920			9087
50	CTC ACC TGC ATG TAC ACA GGA AAG GAG AAA GTC AGT CCG GGT AGA ACT Leu Thr Cys Met Tyr Thr Gly Lys Glu Lys Val Ser Pro Gly Arg Thr 2925 2930 2935 2940			9135
	TCA GAC CCT AAT CCT GCA GCC CCC GAC AGC GAG TCA GTG ATT GTT GCT Ser Asp Pro Asn Pro Ala Ala Pro Asp Ser Glu Ser Val Ile Val Ala 2945 2950 2955			9183
55	ATG GAG CGG GTA TCT GTT CTT TTT GAT AGG ATC AGG AAA GGC TTT CCT Met Glu Arg Val Ser Val Leu Phe Asp Arg Ile Arg Lys Gly Phe Pro			9231

EP 0 614 977 A2

	2960	2965	2970	
5	TGT GAA GCC AGA GTG GTG GCC AGG ATC CTG CCC CAG TTT CTA GAC GAC Cys Glu Ala Arg Val Val Ala Arg Ile Leu Pro Gln Phe Leu Asp Asp 2975 2980 2985			9279
10	TTC TTC CCA CCC CAG GAC ATC ATG AAC AAA GTC ATC GGA GAG TTT CTG Phe Phe Pro Pro Gln Asp Ile Met Asn Lys Val Ile Gly Glu Phe Leu 2990 2995 3000			9327
15	TCC AAC CAG CAG CCA TAC CCC CAG TTC ATG GCC ACC GTG GTG TAT AAG Ser Asn Gln Gln Pro Tyr Pro Gln Phe Met Ala Thr Val Val Tyr Lys 3005 3010 3015 3020			9375
20	GTG TTT CAG ACT CTG CAC AGC ACC GGG CAG TCG TCC ATG GTC CGG GAC Val Phe Gln Thr Leu His Ser Thr Gly Gln Ser Ser Met Val Arg Asp 3025 3030 3035			9423
25	TGG GTC ATG CTG TCC CTC TCC AAC TTC ACG CAG AGG GCC CCG GTC GCC Trp Val Met Leu Ser Leu Ser Asn Phe Thr Gln Arg Ala Pro Val Ala 3040 3045 3050			9471
30	ATG GCC ACG TGG AGC CTC TCC TGC TTC TTT GTC AGC GCG TCC ACC AGC Met Ala Thr Trp Ser Leu Ser Cys Phe Phe Val Ser Ala Ser Thr Ser 3055 3060 3065			9519
35	CCG TGG GTC GCG GCG ATC CTC CCA CAT GTC ATC AGC AGG ATG GGC AAG Pro Trp Val Ala Ala Ile Leu Pro His Val Ile Ser Arg Met Gly Lys 3070 3075 3080			9567
40	CTG GAG CAG GTG GAC GTG AAC CTT TTC TGC CTG GTC GCC ACA GAC TTC Leu Glu Gln Val Asp Val Asn Leu Phe Cys Leu Val Ala Thr Asp Phe 3085 3090 3095 3100			9615
45	TAC AGA CAC CAG ATA GAG GAG GAG CTC GAC CGC AGG GCC TTC CAG TCT Tyr Arg His Gln Ile Glu Glu Glu Leu Asp Arg Arg Ala Phe Gln Ser 3105 3110 3115			9663
50	GTG CTT GAG GTG GTT GCA GCC CCA GGA AGC CCA TAT CAC CGG CTG CTG Val Leu Glu Val Val Ala Ala Pro Gly Ser Pro Tyr His Arg Leu Leu 3120 3125 3130			9711
55	ACT TGT TTA CGA AAT GTC CAC AAG GTC ACC ACC TGC T GAGCGCCATG Thr Cys Leu Arg Asn Val His Lys Val Thr Thr Cys 3135 3140			9758
60	GTGGGAGAGA CTGTGAGGCG GCAGCTGGGG CCGGAGCCTT TGAAGTCTG TGCCCTTGTG			9818
65	CCCTGCCTCC ACCGAGCCAG CTTGGTCCCT ATGGGCTTCC GCACATGCCG CGGGCGGCCA			9878
70	GGCAACGTGC GTGTCTCTGC CATGTGGCAG AAGTGCTCTT TGTGGCAGTG GCCAGGCAGG			9938
75	GAGTGTCTGC AGTCCTGGTG GGGCTGAGCC TGAGGCCTTC CAGAAAGCAG GAGCAGCTGT			9998
80	GCTGCACCCC ATGTGGGTGA CCAGGTCCTT TCTCCTGATA GTCACCTGCT GGTGTGTGCC			10058
85	AGGTTGCAGC TGCTCTTGCA TCTGGGCCAG AAGTCCTCCC TCCTGCAGGC TGGCTGTTGG			10118
90	CCCCCTGCT GTCCTGCAGT AGAAGGTGCC GTGAGCAGGC TTTGGGAACA CTGGCCTGGG			10178
95	TCTCCCTGGT GGGGTGTGCA TGCCACGCCC CGTGTCTGGA TGCACAGATG CCATGGCCTG			10238
100	TGCTGGGCCA GTGGCTGGGG GTGCTAGACA CCCGGCACCA TTCTCCCTTC TCTCTTTTCT			10298
105	TCTCAGGATT TAAAATTTAA TTATATCAGT AAAGAGATTA ATTTTAACGT AAAAAAAAAA			10358
110	AAAAAAAA			10366

55

(2) INFORMATION FOR SEQ ID NO:6:

(i) SEQUENCE CHARACTERISTICS:

EP 0 614 977 A2

(A) LENGTH: 3144 amino acids
(B) TYPE: amino acid
(D) TOPOLOGY: linear

(ii) MOLECULE TYPE: protein

5

(xi) SEQUENCE DESCRIPTION: SEQ ID NO:5:

Met Ala Thr Leu Glu Lys Leu Met Lys Ala Phe Glu Ser Leu Lys Ser
1 5 10 15

Phe Gln Gln Gln Gln Gln Gln Gln Gln Gln Gln Gln Gln Gln Gln
20 25 30

Gln Gln Gln Gln Gln Gln Gln Gln Gln Pro Pro Pro Pro Pro Pro Pro
35 40 45

Pro Pro Pro Gln Leu Pro Gln Pro Pro Pro Gln Ala Gln Pro Leu Leu
50 55 60

Pro Gln Pro Gln Pro Pro Pro Pro Pro Pro Pro Pro Gly Pro
65 70 75 80

Ala Val Ala Glu Glu Pro Leu His Arg Pro Lys Lys Glu Leu Ser Ala
85 90 95

Thr Lys Lys Asp Arg Val Asn His Cys Leu Thr Ile Cys Glu Asn Ile
100 105 110

Val Ala Gln Ser Val Arg Asn Ser Pro Glu Phe Gln Lys Leu Leu Gly
115 120 125

Ile Ala Met Glu Leu Phe Leu Leu Cys Ser Asp Asp Ala Glu Ser Asp
130 135 140

Val Arg Met Val Ala Asp Glu Cys Leu Asn Lys Val Ile Lys Ala Leu
145 150 155 160

Met Asp Ser Asn Leu Pro Arg Leu Gln Leu Glu Leu Tyr Lys Glu Ile
165 170 175

Lys Lys Asn Gly Ala Pro Arg Ser Leu Arg Ala Ala Leu Trp Arg Phe
180 185 190

Ala Glu Leu Ala His Leu Val Arg Pro Gln Lys Cys Arg Pro Tyr Leu
195 200 205

Val Asn Leu Leu Pro Cys Leu Thr Arg Thr Ser Lys Arg Pro Glu Glu
210 215 220

Ser Val Gln Glu Thr Leu Ala Ala Ala Val Pro Lys Ile Met Ala Ser
225 230 235 240

Phe Gly Asn Phe Ala Asn Asp Asn Glu Ile Lys Val Leu Leu Lys Ala
245 250 255

Phe Ile Ala Asn Leu Lys Ser Ser Ser Pro Thr Ile Arg Arg Thr Ala
260 265 270

Ala Gly Ser Ala Val Ser Ile Cys Gln His Ser Arg Arg Thr Gln Tyr
275 280 285

Phe Tyr Ser Trp Leu Leu Asn Val Leu Leu Gly Leu Leu Val Pro Val
290 295 300

Glu Asp Glu His Ser Thr Leu Leu Ile Leu Gly Val Leu Leu Thr Leu
305 310 315 320

Arg Tyr Leu Val Pro Leu Leu Gln Gln Val Lys Asp Thr Ser Leu
325 330 335

Lys Gly Ser Phe Gly Val Thr Arg Lys Glu Met Glu Val Ser Pro Ser

EP 0 614 977 A2

		340		345		350	
		Ala Glu Gln Leu Val Gln Val Tyr Glu Leu Thr Leu His His Thr Gln					
		355		360		365	
5		His Gln Asp His Asn Val Val Thr Gly Ala Leu Glu Leu Leu Gln Gln					
		370		375		380	
		Leu Phe Arg Thr Pro Pro Pro Glu Leu Leu Gln Thr Leu Thr Ala Val					
		385		390		395	400
10		Gly Gly Ile Gly Gln Leu Thr Ala Ala Lys Glu Glu Ser Gly Gly Arg					
			405		410		415
		Ser Arg Ser Gly Ser Ile Val Glu Leu Ile Ala Gly Gly Gly Ser Ser					
			420		425		430
15		Cys Ser Pro Val Leu Ser Arg Lys Gln Lys Gly Lys Val Leu Leu Gly					
			435		440		445
		Glu Glu Glu Ala Leu Glu Asp Asp Ser Glu Ser Arg Ser Asp Val Ser					
			450		455		460
20		Ser Ser Ala Leu Thr Ala Ser Val Lys Asp Glu Ile Ser Gly Glu Leu					
				470		475	480
		Ala Ala Ser Ser Gly Val Ser Thr Pro Gly Ser Ala Gly His Asp Ile					
			485		490		495
25		Ile Thr Glu Gln Pro Arg Ser Gln His Thr Leu Gln Ala Asp Ser Leu					
			500		505		510
		Asp Leu Ala Ser Cys Asp Leu Thr Ser Ser Ala Thr Asp Gly Asp Glu					
			515		520		525
30		Glu Asp Ile Leu Ser His Ser Ser Ser Gln Val Ser Ala Val Pro Ser					
			530		535		540
		Asp Pro Ala Met Asp Leu Asn Asp Gly Thr Gln Ala Ser Ser Pro Ile					
			545		550		555
		Ser Asp Ser Ser Gln Thr Thr Thr Glu Gly Pro Asp Ser Ala Val Thr					
			565		570		575
35		Pro Ser Asp Ser Ser Glu Ile Val Leu Asp Gly Thr Asp Asn Gln Tyr					
			580		585		590
		Leu Gly Leu Gln Ile Gly Gln Pro Gln Asp Glu Asp Glu Glu Ala Thr					
			595		600		605
40		Gly Ile Leu Pro Asp Glu Ala Ser Glu Ala Phe Arg Asn Ser Ser Met					
			610		615		620
		Ala Leu Gln Gln Ala His Leu Leu Lys Asn Met Ser His Cys Arg Gln					
			625		630		635
45		Pro Ser Asp Ser Ser Val Asp Lys Phe Val Leu Arg Asp Glu Ala Thr					
			645		650		655
		Glu Pro Gly Asp Gln Glu Asn Lys Pro Cys Arg Ile Lys Gly Asp Ile					
			660		665		670
50		Gly Gln Ser Thr Asp Asp Asp Ser Ala Pro Leu Val His Ser Val Arg					
			675		680		685
		Leu Leu Ser Ala Ser Phe Leu Leu Thr Gly Gly Lys Asn Val Leu Val					
			690		695		700
55		Pro Asp Arg Asp Val Arg Val Ser Val Lys Ala Leu Ala Leu Ser Cys					
			705		710		715
		Val Gly Ala Ala Val Ala Leu His Pro Glu Ser Phe Phe Ser Lys Leu					
			725		730		735

EP 0 614 977 A2

Tyr Lys Val Pro Leu Asp Thr Thr Glu Tyr Pro Glu Glu Gln Tyr Val
 740 745 750
 5 Ser Asp Ile Leu Asn Tyr Ile Asp His Gly Asp Pro Gln Val Arg Gly
 755 760 765
 Ala Thr Ala Ile Leu Cys Gly Thr Leu Ile Cys Ser Ile Leu Ser Arg
 770 775 780
 10 Ser Arg Phe His Val Gly Asp Trp Met Gly Thr Ile Arg Thr Leu Thr
 785 790 795 800
 Gly Asn Thr Phe Ser Leu Ala Asp Cys Ile Pro Leu Leu Arg Lys Thr
 805 810 815
 15 Leu Lys Asp Glu Ser Ser Val Thr Cys Lys Leu Ala Cys Thr Ala Val
 820 825 830
 Arg Asn Cys Val Met Ser Leu Cys Ser Ser Ser Tyr Ser Glu Leu Gly
 835 840 845
 20 Leu Gln Leu Ile Ile Asp Val Leu Thr Leu Arg Asn Ser Ser Tyr Trp
 850 855 860
 Leu Val Arg Thr Glu Leu Leu Glu Thr Leu Ala Glu Ile Asp Phe Arg
 865 870 875 880
 25 Leu Val Ser Phe Leu Glu Ala Lys Ala Glu Asn Leu His Arg Gly Ala
 885 890 895
 His His Tyr Thr Gly Leu Leu Lys Leu Gln Glu Arg Val Leu Asn Asn
 900 905 910
 Val Val Ile His Leu Leu Gly Asp Glu Asp Pro Arg Val Arg His Val
 915 920 925
 30 Ala Ala Ala Ser Leu Ile Arg Leu Val Pro Lys Leu Phe Tyr Lys Cys
 930 935 940
 Asp Gln Gly Gln Ala Asp Pro Val Val Ala Val Ala Arg Asp Gln Ser
 945 950 955 960
 35 Ser Val Tyr Leu Lys Leu Leu Met His Glu Thr Gln Pro Pro Ser His
 965 970 975
 Phe Ser Val Ser Thr Ile Thr Arg Ile Tyr Arg Gly Tyr Asn Leu Leu
 980 985 990
 40 Pro Ser Ile Thr Asp Val Thr Met Glu Asn Asn Leu Ser Arg Val Ile
 995 1000 1005
 Ala Ala Val Ser His Glu Leu Ile Thr Ser Thr Thr Arg Ala Leu Thr
 1010 1015 1020
 45 Phe Gly Cys Cys Glu Ala Leu Cys Leu Leu Ser Thr Ala Phe Pro Val
 1025 1030 1035 1040
 Cys Ile Trp Ser Leu Gly Trp His Cys Gly Val Pro Pro Leu Ser Ala
 1045 1050 1055
 50 Ser Asp Glu Ser Arg Lys Ser Cys Thr Val Gly Met Ala Thr Met Ile
 1060 1065 1070
 Leu Thr Leu Leu Ser Ser Ala Trp Phe Pro Leu Asp Leu Ser Ala His
 1075 1080 1085
 55 Gln Asp Ala Leu Ile Leu Ala Gly Asn Leu Leu Ala Ala Ser Ala Pro
 1090 1095 1100
 Lys Ser Leu Arg Ser Ser Trp Ala Ser Glu Glu Glu Ala Asn Pro Ala
 1105 1110 1115 1120

EP 0 614 977 A2

Ala Thr Lys Gln Glu Glu Val Trp Pro Ala Leu Gly Asp Arg Ala Leu
1125 1130 1135

5 Val Pro Met Val Glu Gln Leu Phe Ser His Leu Leu Lys Val Ile Asn
1140 1145 1150

Ile Cys Ala His Val Leu Asp Asp Val Ala Pro Gly Pro Ala Ile Lys
1155 1160 1165

10 Ala Ala Leu Pro Ser Leu Thr Asn Pro Pro Ser Leu Ser Pro Ile Arg
1170 1175 1180

Arg Lys Gly Lys Glu Lys Glu Pro Gly Glu Gln Ala Ser Val Pro Leu
1185 1190 1195 1200

15 Ser Pro Lys Lys Gly Ser Glu Ala Ser Ala Ala Ser Arg Gln Ser Asp
1205 1210 1215

Thr Ser Gly Pro Val Thr Thr Ser Lys Ser Ser Ser Leu Gly Ser Phe
1220 1225 1230

20 Tyr His Leu Pro Ser Tyr Leu Arg Leu His Asp Val Leu Lys Ala Thr
1235 1240 1245

His Ala Asn Tyr Lys Val Thr Leu Asp Leu Gln Asn Ser Thr Glu Lys
1250 1255 1260

25 Phe Gly Gly Phe Leu Arg Ser Ala Leu Asp Val Leu Ser Gln Ile Leu
1265 1270 1275 1280

Glu Leu Ala Thr Leu Gln Asp Ile Gly Lys Cys Val Glu Glu Ile Leu
1285 1290 1295

30 Gly Tyr Leu Lys Ser Cys Phe Ser Arg Glu Pro Met Met Ala Thr Val
1300 1305 1310

Cys Val Gln Gln Leu Leu Lys Thr Leu Phe Gly Thr Asn Leu Ala Ser
1315 1320 1325

Gln Phe Asp Gly Leu Ser Ser Asn Pro Ser Lys Ser Gln Gly Arg Ala
1330 1335 1340

35 Gln Arg Leu Gly Ser Ser Ser Val Arg Pro Gly Leu Tyr His Tyr Cys
1345 1350 1355 1360

Phe Met Ala Pro Tyr Thr His Phe Thr Gln Ala Leu Ala Asp Ala Ser
1365 1370 1375

40 Leu Arg Asn Met Val Gln Ala Glu Gln Glu Asn Asp Thr Ser Gly Trp
1380 1385 1390

Phe Asp Val Leu Gln Lys Val Ser Thr Gln Leu Lys Thr Asn Leu Thr
1395 1400 1405

45 Ser Val Thr Lys Asn Arg Ala Asp Lys Asn Ala Ile His Asn His Ile
1410 1415 1420

Arg Leu Phe Glu Pro Leu Val Ile Lys Ala Leu Lys Gln Tyr Thr Thr
1425 1430 1435 1440

50 Thr Thr Cys Val Gln Leu Gln Lys Gln Val Leu Asp Leu Leu Ala Gln
1445 1450 1455

Leu Val Gln Leu Arg Val Asn Tyr Cys Leu Leu Asp Ser Asp Gln Val
1460 1465 1470

55 Phe Ile Gly Phe Val Leu Lys Gln Phe Glu Tyr Ile Glu Val Gly Gln
1475 1480 1485

Phe Arg Glu Ser Glu Ala Ile Ile Pro Asn Ile Phe Phe Phe Leu Val
1490 1495 1500

EP 0 614 977 A2

Leu Leu Ser Tyr Glu Arg Tyr His Ser Lys Gln Ile Ile Gly Ile Pro
 1505 1510 1515 1520
 Lys Ile Ile Gln Leu Cys Asp Gly Ile Met Ala Ser Gly Arg Lys Ala
 1525 1530 1535
 5
 Val Thr His Ala Ile Pro Ala Leu Gln Pro Ile Val His Asp Leu Phe
 1540 1545 1550
 Val Leu Arg Gly Thr Asn Lys Ala Asp Ala Gly Lys Glu Leu Glu Thr
 1555 1560 1565
 10
 Gln Lys Glu Val Val Val Ser Met Leu Leu Arg Leu Ile Gln Tyr His
 1570 1575 1580
 Gln Val Leu Glu Met Phe Ile Leu Val Leu Gln Gln Cys His Lys Glu
 1585 1590 1595 1600
 15
 Asn Glu Asp Lys Trp Lys Arg Leu Ser Arg Gln Ile Ala Asp Ile Ile
 1605 1610 1615
 Leu Pro Met Leu Ala Lys Gln Gln Met His Ile Asp Ser His Glu Ala
 1620 1625 1630
 20
 Leu Gly Val Leu Asn Thr Leu Phe Glu Ile Leu Ala Pro Ser Ser Leu
 1635 1640 1645
 Arg Pro Val Asp Met Leu Leu Arg Ser Met Phe Val Thr Pro Asn Thr
 1650 1655 1660
 25
 Met Ala Ser Val Ser Thr Val Gln Leu Trp Ile Ser Gly Ile Leu Ala
 1665 1670 1675 1680
 Ile Leu Arg Val Leu Ile Ser Gln Ser Thr Glu Asp Ile Val Leu Ser
 1685 1690 1695
 30
 Arg Ile Gln Glu Leu Ser Phe Ser Pro Tyr Leu Ile Ser Cys Thr Val
 1700 1705 1710
 Ile Asn Arg Leu Arg Asp Gly Asp Ser Thr Ser Thr Leu Glu Glu His
 1715 1720 1725
 35
 Ser Glu Gly Lys Gln Ile Lys Asn Leu Pro Glu Glu Thr Phe Ser Arg
 1730 1735 1740
 Phe Leu Leu Gln Leu Val Gly Ile Leu Leu Glu Asp Ile Val Thr Lys
 1745 1750 1755 1760
 40
 Gln Leu Lys Val Glu Met Ser Glu Gln Gln His Thr Phe Tyr Cys Gln
 1765 1770 1775
 Glu Leu Gly Thr Leu Leu Met Cys Leu Ile His Ile Phe Lys Ser Gly
 1780 1785 1790
 45
 Met Phe Arg Arg Ile Thr Ala Ala Ala Thr Arg Leu Phe Arg Ser Asp
 1795 1800 1805
 Gly Cys Gly Gly Ser Phe Tyr Thr Leu Asp Ser Leu Asn Leu Arg Ala
 1810 1815 1820
 50
 Arg Ser Met Ile Thr Thr His Pro Ala Leu Val Leu Leu Trp Cys Gln
 1825 1830 1835 1840
 Ile Leu Leu Leu Val Asn His Thr Asp Tyr Arg Trp Trp Ala Glu Val
 1845 1850 1855
 55
 Gln Gln Thr Pro Lys Arg His Ser Leu Ser Ser Thr Lys Leu Leu Ser
 1860 1865 1870
 Pro Gln Met Ser Gly Glu Glu Glu Asp Ser Asp Leu Ala Ala Lys Leu
 1875 1880 1885

EP 0 614 977 A2

Gly Met Cys Asn Arg Glu Ile Val Arg Arg Gly Ala Leu Ile Leu Phe
 1890 1895 1900
 Cys Asp Tyr Val Cys Gln Asn Leu His Asp Ser Glu His Leu Thr Trp
 1905 1910 1915 1920
 5
 Leu Ile Val Asn His Ile Gln Asp Leu Ile Ser Leu Ser His Glu Pro
 1925 1930 1935
 Pro Val Gln Asp Phe Ile Ser Ala Val His Arg Asn Ser Ala Ala Ser
 1940 1945 1950
 10
 Gly Leu Phe Ile Gln Ala Ile Gln Ser Arg Cys Glu Asn Leu Ser Thr
 1955 1960 1965
 Pro Thr Met Leu Lys Lys Thr Leu Gln Cys Leu Glu Gly Ile His Leu
 1970 1975 1980
 15
 Ser Gln Ser Gly Ala Val Leu Thr Leu Tyr Val Asp Arg Leu Leu Cys
 1985 1990 1995 2000
 Thr Pro Phe Arg Val Leu Ala Arg Met Val Asp Ile Leu Ala Cys Arg
 2005 2010 2015
 20
 Arg Val Glu Met Leu Leu Ala Ala Asn Leu Gln Ser Ser Met Ala Gln
 2020 2025 2030
 Leu Pro Met Glu Glu Leu Asn Arg Ile Gln Glu Tyr Leu Gln Ser Ser
 2035 2040 2045
 25
 Gly Leu Ala Gln Arg His Gln Arg Leu Tyr Ser Leu Leu Asp Arg Phe
 2050 2055 2060
 Arg Leu Ser Thr Met Gln Asp Ser Leu Ser Pro Ser Pro Pro Val Ser
 2065 2070 2075 2080
 30
 Ser His Pro Leu Asp Gly Asp Gly His Val Ser Leu Glu Thr Val Ser
 2085 2090 2095
 Pro Asp Lys Asp Trp Tyr Val His Leu Val Lys Ser Gln Cys Trp Thr
 2100 2105 2110
 35
 Arg Ser Asp Ser Ala Leu Leu Glu Gly Ala Glu Leu Val Asn Arg Ile
 2115 2120 2125
 Pro Ala Glu Asp Met Asn Ala Phe Met Met Asn Ser Glu Phe Asn Leu
 2130 2135 2140
 40
 Ser Leu Leu Ala Pro Cys Leu Ser Leu Gly Met Ser Glu Ile Ser Gly
 2145 2150 2155 2160
 Gly Gln Lys Ser Ala Leu Phe Glu Ala Ala Arg Glu Val Thr Leu Ala
 2165 2170 2175
 45
 Arg Val Ser Gly Thr Val Gln Gln Leu Pro Ala Val His His Val Phe
 2180 2185 2190
 Gln Pro Glu Leu Pro Ala Glu Pro Ala Ala Tyr Trp Ser Lys Leu Asn
 2195 2200 2205
 50
 Asp Leu Phe Gly Asp Ala Ala Leu Tyr Gln Ser Leu Pro Thr Leu Ala
 2210 2215 2220
 Arg Ala Leu Ala Gln Tyr Leu Val Val Val Ser Lys Leu Pro Ser His
 2225 2230 2235 2240
 Leu His Leu Pro Pro Glu Lys Glu Lys Asp Ile Val Lys Phe Val Val
 2245 2250 2255
 55
 Ala Thr Leu Glu Ala Leu Ser Trp His Leu Ile His Glu Gln Ile Pro
 2260 2265 2270

EP 0 614 977 A2

Leu Ser Leu Asp Leu Gln Ala Gly Leu Asp Cys Cys Cys Leu Ala Leu
 2275 2280 2285
 5 Gln Leu Pro Gly Leu Trp Ser Val Val Ser Ser Thr Glu Phe Val Thr
 2290 2295 2300
 His Ala Cys Ser Leu Ile Tyr Cys Val His Phe Ile Leu Glu Ala Val
 2305 2310 2315 2320
 10 Ala Val Gln Pro Gly Glu Gln Leu Leu Ser Pro Glu Arg Arg Thr Asn
 2325 2330 2335
 Thr Pro Lys Ala Ile Ser Glu Glu Glu Glu Glu Val Asp Pro Asn Thr
 2340 2345 2350
 15 Gln Asn Pro Lys Tyr Ile Thr Ala Ala Cys Glu Met Val Ala Glu Met
 2355 2360 2365
 Val Glu Ser Leu Gln Ser Val Leu Ala Leu Gly His Lys Arg Asn Ser
 2370 2375 2380
 20 Gly Val Pro Ala Phe Leu Thr Pro Leu Leu Arg Asn Ile Ile Ile Ser
 2385 2390 2395 2400
 Leu Ala Arg Leu Pro Leu Val Asn Ser Tyr Thr Arg Val Pro Pro Leu
 2405 2410 2415
 Val Trp Lys Leu Gly Trp Ser Pro Lys Pro Gly Gly Asp Phe Gly Thr
 2420 2425 2430
 25 Ala Phe Pro Glu Ile Pro Val Glu Phe Leu Gln Glu Lys Glu Val Phe
 2435 2440 2445
 Lys Glu Phe Ile Tyr Arg Ile Asn Thr Leu Gly Trp Thr Ser Arg Thr
 2450 2455 2460
 30 Gln Phe Glu Glu Thr Trp Ala Thr Leu Leu Gly Val Leu Val Thr Gln
 2465 2470 2475 2480
 Pro Leu Val Met Glu Gln Glu Glu Ser Pro Pro Glu Glu Asp Thr Glu
 2485 2490 2495
 35 Arg Thr Gln Ile Asn Val Leu Ala Val Gln Ala Ile Thr Ser Leu Val
 2500 2505 2510
 Leu Ser Ala Met Thr Val Pro Val Ala Gly Asn Pro Ala Val Ser Cys
 2515 2520 2525
 40 Leu Glu Gln Gln Pro Arg Asn Lys Pro Leu Lys Ala Leu Asp Thr Arg
 2530 2535 2540
 Phe Gly Arg Lys Leu Ser Ile Ile Arg Gly Ile Val Glu Gln Glu Ile
 2545 2550 2555 2560
 45 Gln Ala Met Val Ser Lys Arg Glu Asn Ile Ala Thr His His Leu Tyr
 2565 2570 2575
 Gln Ala Trp Asp Pro Val Pro Ser Leu Ser Pro Ala Thr Thr Gly Ala
 2580 2585 2590
 50 Leu Ile Ser His Glu Lys Leu Leu Leu Gln Ile Asn Pro Glu Arg Glu
 2595 2600 2605
 Leu Gly Ser Met Ser Tyr Lys Leu Gly Gln Val Ser Ile His Ser Val
 2610 2615 2620
 55 Trp Leu Gly Asn Ser Ile Thr Pro Leu Arg Glu Glu Glu Trp Asp Glu
 2625 2630 2635 2640
 Glu Glu Glu Glu Glu Ala Asp Ala Pro Ala Pro Ser Ser Pro Pro Thr
 2645 2650 2655

EP 0 614 977 A2

Ser Pro Val Asn Ser Arg Lys His Arg Ala Gly Val Asp Ile His Ser
 2660 2665 2670
 Cys Ser Gln Phe Leu Leu Glu Leu Tyr Ser Arg Trp Ile Leu Pro Ser
 2675 2680 2685
 Ser Ser Ala Arg Arg Thr Pro Ala Ile Leu Ile Ser Glu Val Val Arg
 2690 2695 2700
 Ser Leu Leu Val Val Ser Asp Leu Phe Thr Glu Arg Asn Gln Phe Glu
 2705 2710 2715 2720
 Leu Met Tyr Val Thr Leu Thr Glu Leu Arg Arg Val His Pro Ser Glu
 2725 2730 2735
 Asp Glu Ile Leu Ala Gln Tyr Leu Val Pro Ala Thr Cys Lys Ala Ala
 2740 2745 2750
 Ala Val Leu Gly Met Asp Lys Ala Val Ala Glu Pro Val Ser Arg Leu
 2755 2760 2765
 Leu Glu Ser Thr Leu Arg Ser Ser His Leu Pro Ser Arg Val Gly Ala
 2770 2775 2780
 Leu His Gly Ile Leu Tyr Val Leu Glu Cys Asp Leu Leu Asp Asp Thr
 2785 2790 2795 2800
 Ala Lys Gln Leu Ile Pro Val Ile Ser Asp Tyr Leu Leu Ser Asn Leu
 2805 2810 2815
 Lys Gly Ile Ala His Cys Val Asn Ile His Ser Gln Gln His Val Leu
 2820 2825 2830
 Val Met Cys Ala Thr Ala Phe Tyr Leu Ile Glu Asn Tyr Pro Leu Asp
 2835 2840 2845
 Val Gly Pro Glu Phe Ser Ala Ser Ile Ile Gln Met Cys Gly Val Met
 2850 2855 2860
 Leu Ser Gly Ser Glu Glu Ser Thr Pro Ser Ile Ile Tyr His Cys Ala
 2865 2870 2875 2880
 Leu Arg Gly Leu Glu Arg Leu Leu Leu Ser Glu Gln Leu Ser Arg Leu
 2885 2890 2895
 Asp Ala Glu Ser Leu Val Lys Leu Ser Val Asp Arg Val Asn Val His
 2900 2905 2910
 Ser Pro His Arg Ala Met Ala Ala Leu Gly Leu Met Leu Thr Cys Met
 2915 2920 2925
 Tyr Thr Gly Lys Glu Lys Val Ser Pro Gly Arg Thr Ser Asp Pro Asn
 2930 2935 2940
 Pro Ala Ala Pro Asp Ser Glu Ser Val Ile Val Ala Met Glu Arg Val
 2945 2950 2955 2960
 Ser Val Leu Phe Asp Arg Ile Arg Lys Gly Phe Pro Cys Glu Ala Arg
 2965 2970 2975
 Val Val Ala Arg Ile Leu Pro Gln Phe Leu Asp Asp Phe Phe Pro Pro
 2980 2985 2990
 Gln Asp Ile Met Asn Lys Val Ile Gly Glu Phe Leu Ser Asn Gln Gln
 2995 3000 3005
 Pro Tyr Pro Gln Phe Met Ala Thr Val Val Tyr Lys Val Phe Gln Thr
 3010 3015 3020
 Leu His Ser Thr Gly Gln Ser Ser Met Val Arg Asp Trp Val Met Leu
 3025 3030 3035 3040

EP 0 614 977 A2

Ser Leu Ser Asn Phe Thr Gln Arg Ala Pro Val Ala Met Ala Thr Trp
 3045 3050 3055

Ser Leu Ser Cys Phe Phe Val Ser Ala Ser Thr Ser Pro Trp Val Ala
 3060 3065 3070

5

Ala Ile Leu Pro His Val Ile Ser Arg Met Gly Lys Leu Glu Gln Val
 3075 3080 3085

Asp Val Asn Leu Phe Cys Leu Val Ala Thr Asp Phe Tyr Arg His Gln
 3090 3095 3100

10

Ile Glu Glu Glu Leu Asp Arg Arg Ala Phe Gln Ser Val Leu Glu Val
 3105 3110 3115 3120

Val Ala Ala Pro Gly Ser Pro Tyr His Arg Leu Leu Thr Cys Leu Arg
 3125 3130 3135

15

Asn Val His Lys Val Thr Thr Cys
 3140

20

25

30

35

40

45

50

55

Claims

- 5 1. An isolated, purified or recombinant polypeptide comprising a huntingtin protein or a mutant, fragment or variant thereof having substantially the same activity as huntingtin protein.
2. A polypeptide according to claim 1 having the amino acid sequence shown in SEQ ID NO:6.
- 10 3. A polypeptide according to claim 1 or 2 which is essentially purified and/or has at least 5 contiguous amino acids.
4. An isolated, purified or recombinant nucleic acid molecule comprising nucleic acid which is:
 - 15 (a) a sequence encoding a huntingtin protein according to any preceding claim (whether normal or genetically defective), or its complementary strand;
 - (b) a sequence that is substantially homologous to, or hybridises under stringent conditions to, either sequence in (a);
 - (c) a sequence that is substantially homologous to, or would hybridise under stringent conditions to, a sequence in (a) or (b) but for the degeneracy of the genetic code;
 - 20 or a fragment of any of (a), (b) or (c).
5. A nucleic acid according to claim 1, wherein the huntingtin protein has the amino acid sequence shown in SEQ ID NO:6 and/or the nucleic acid is DNA encoding the amino acid sequence SEQ ID NO:5.
- 25 6. A nucleic acid molecule according to claim 4 or 5 which is a probe for detecting the presence of huntingtin in a sample comprising being at least 5, such as at least 15, contiguous nucleotides.
7. A (preferably recombinant) nucleic acid molecule according to any of claims 4 to 6 comprising a transcriptional region functional in a cell operably linked to a sequence complimentary to an RNA sequence encoding a protein according to any of claims 1 to 3 or at least 5 contiguous amino acids thereof.
- 30 8. A vector comprising a nucleic acid molecule according to any of claims 4 to 7.
9. A vector according to claim 8 wherein the nucleic acid molecule, such as encoding huntingtin protein, is operably linked to transcriptional and/or translational expression signals.
- 35 10. A host cell transformed or transfected with a vector according to claim 4 or 5.
11. An antibody specific for huntingtin protein, or a protein according to any of claims 1 to 3.
- 40 12. A hybridoma which produces an antibody according to claim 11.
13. A method of detecting the presence of, or predisposition to develop, Huntington's disease in a subject, the method comprising evaluating the characteristics of huntingtin nucleic acid in a sample from the subject, for example in relation to the number of (CAG) repeats.
- 45 14. A method according to claim 13 comprising:
 - (a) taking a sample from the subject;
 - (b) evaluating the characteristics of huntingtin nucleic acid in the sample, wherein the evaluation comprises detecting the huntingtin (CAG)_n region in the sample; and
 - (c) comparing the characteristics found in (b) with a similar analysis from an individual not having, or
 - 50 not suspected of having, Huntington's disease; and
 - (d) the presence of, or predisposition to develop, Huntington's disease being indicated if those characteristics in the huntingtin (CAG)_n region differ.
- 55 15. A method according to claim 13 comprising:
 - (a) taking a sample from a subject and;
 - (b) evaluating the characteristics of huntingtin nucleic acid comprising the huntingtin (CAG)_n region in the sample by Southern blot, northern blot, or polymerase chain reaction analysis.

16. The use of:
- (a) a nucleic acid molecule according to any of claims 4 to 6 or a vector according to claim 8 which encodes a functional (or non-defective) protein;
 - (b) a polypeptide according to any of claims 1 to 3 which is functional (or non-defective);
 - 5 (c) a host cell according to claim 10 expressing a polypeptide which is functional (or non-defective); and/or
 - (d) an antagonist to, or a compound that binds to, huntingdon protein;
 - in the preparation of an agent for treating, delaying or preventing a neurodegenerative disorder.
17. The use according to claim 16 which is gene therapy.
18. The use according to claim 16 or 17 for treating, preventing or delaying Huntingdon's disease.
19. The use according to any of claims 16 to 17 wherein the nucleic acid has from 11 to 34 (CAG) repeats and/or the polypeptide has from 11 to 34 Gln repeats, said repeats being consecutive.
20. A diagnostic and/or immunoassay kit comprising at least one container and;
- (a) a nucleic acid molecule according to any of claims 4 to 6, optionally labelled; or
 - (b) an antibody according to claim 11, optionally labelled.
21. The use of:
- (a) a nucleic acid molecule according to any of claims 4 to 6 or a vector according to claim 8 which encodes a functional (or non-defective) protein;
 - (b) a polypeptide according to any of claims 1 to 3 which is functional (or non-defective);
 - 25 (c) a host cell according to claim 10 expressing a polypeptide which is functional (or non-defective); and/or
 - (d) an antagonist to, or a compound that binds to, huntingdon protein;
 - in the preparation of a medicament.
22. A pharmaceutical composition comprising:
- 30 (a) a nucleic acid molecule according to any of claims 4 to 6 or a vector according to claim 8 which encodes a functional (or non-defective) protein;
 - (b) a polypeptide according to any of claims 1 to 3 which is functional (or non-defective);
 - (c) a host cell according to claim 10 expressing a polypeptide which is functional (or non-defective); and/or
 - 35 (d) an antagonist to, or a compound that binds to, huntingdon protein;
 - in admixture with pharmaceutically acceptable carrier.
23. A process for the preparation of a polypeptide, the process comprising culturing a host cell according to claim 10 under conditions whereby the polypeptide is expressed, and purifying or isolating the polypeptide.

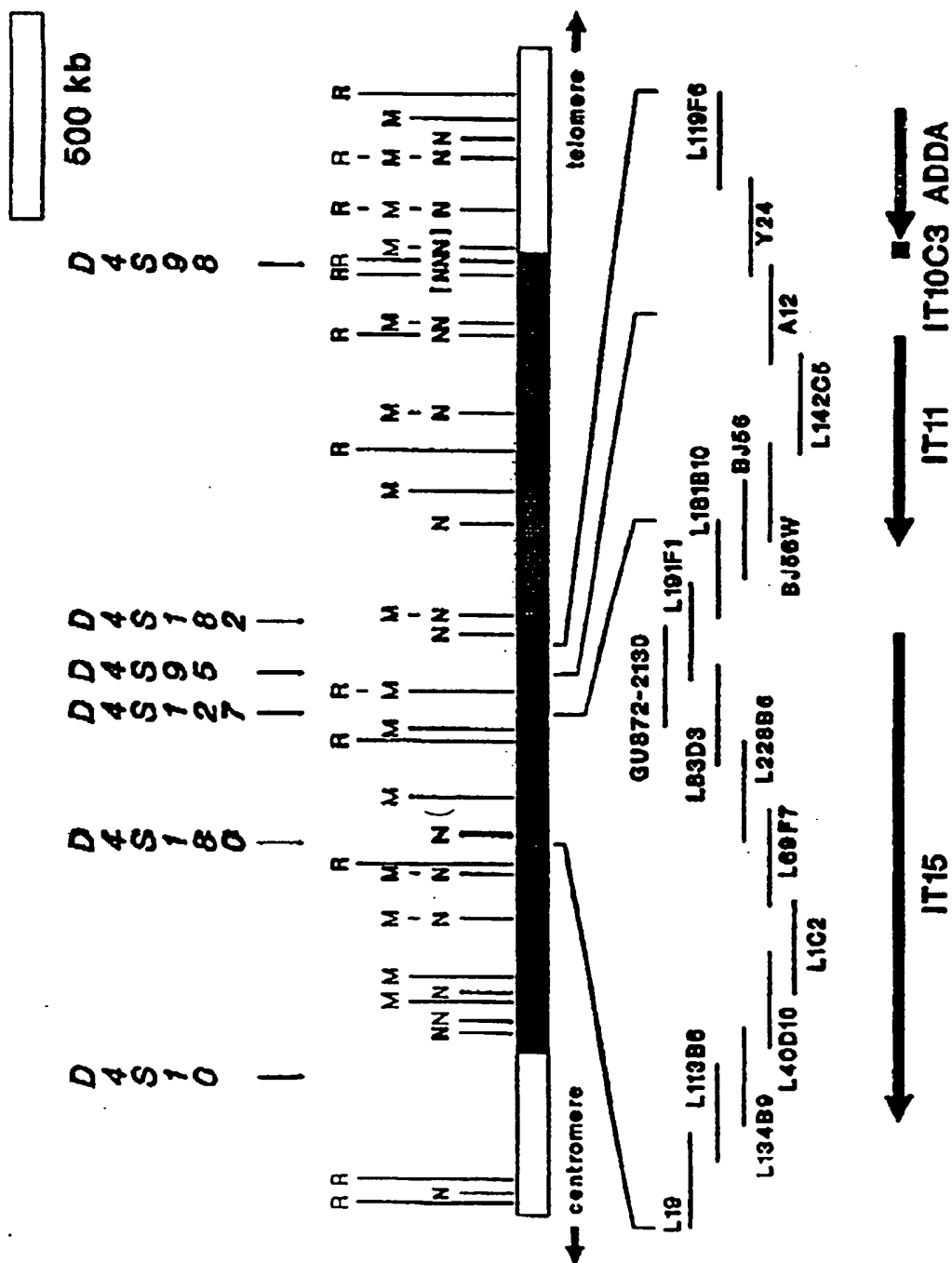


FIGURE 1

1 2 3



— 28 S



— 18 S

Figure 2

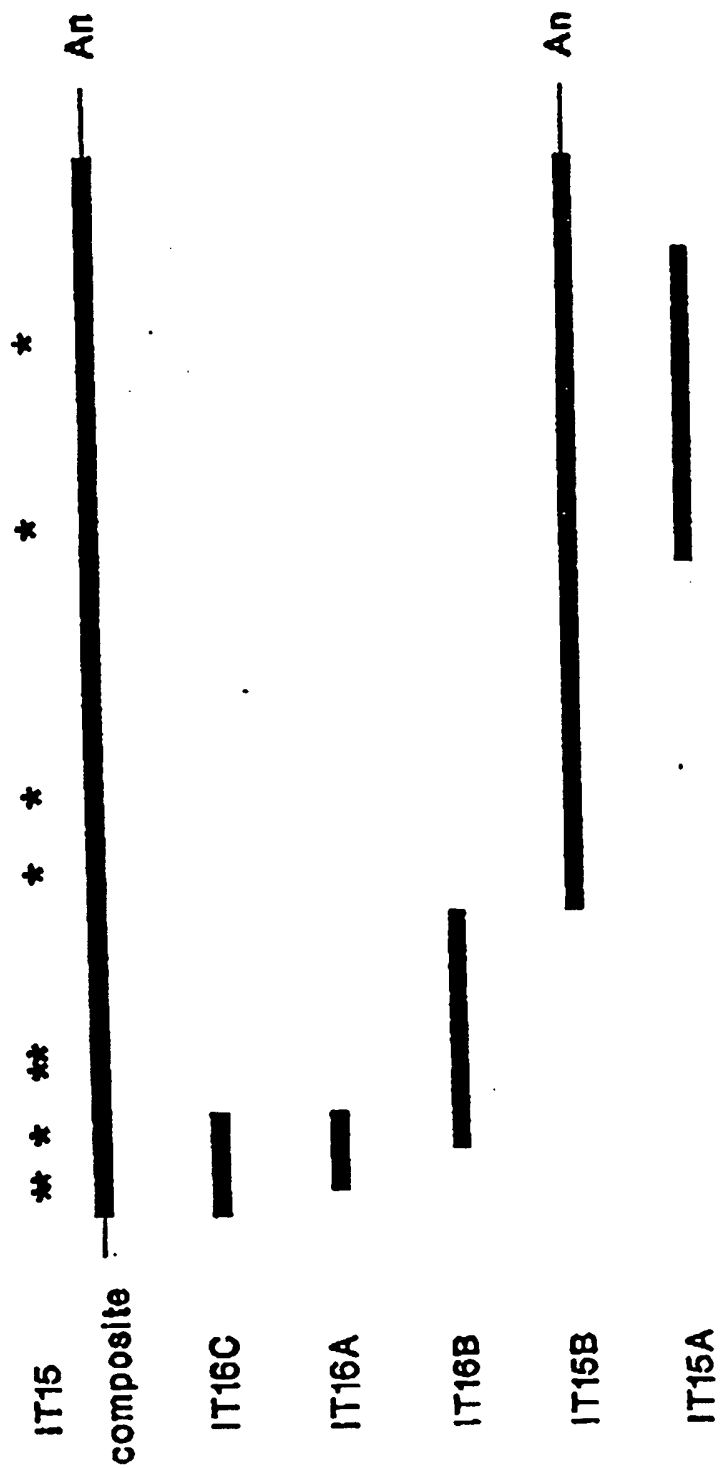


Figure 3.

[illegible]

FIGURE 4 (Sheet 1 of 3)

10541
10542
10543
10544
10545
10546
10547
10548
10549
10550
10551
10552
10553
10554
10555
10556
10557
10558
10559
10560
10561
10562
10563
10564
10565
10566
10567
10568
10569
10570
10571
10572
10573
10574
10575
10576
10577
10578
10579
10580
10581
10582
10583
10584
10585
10586
10587
10588
10589
10590
10591
10592
10593
10594
10595
10596
10597
10598
10599
10600
10601
10602
10603
10604
10605
10606
10607
10608
10609
10610
10611
10612
10613
10614
10615
10616
10617
10618
10619
10620
10621
10622
10623
10624
10625
10626
10627
10628
10629
10630
10631
10632
10633
10634
10635
10636
10637
10638
10639
10640
10641
10642
10643
10644
10645
10646
10647
10648
10649
10650
10651
10652
10653
10654
10655
10656
10657
10658
10659
10660
10661
10662
10663
10664
10665
10666
10667
10668
10669
10670
10671
10672
10673
10674
10675
10676
10677
10678
10679
10680
10681
10682
10683
10684
10685
10686
10687
10688
10689
10690
10691
10692
10693
10694
10695
10696
10697
10698
10699
10700
10701
10702
10703
10704
10705
10706
10707
10708
10709
10710
10711
10712
10713
10714
10715
10716
10717
10718
10719
10720
10721
10722
10723
10724
10725
10726
10727
10728
10729
10730
10731
10732
10733
10734
10735
10736
10737
10738
10739
10740
10741
10742
10743
10744
10745
10746
10747
10748
10749
10750
10751
10752
10753
10754
10755
10756
10757
10758
10759
10760
10761
10762
10763
10764
10765
10766
10767
10768
10769
10770
10771
10772
10773
10774
10775
10776
10777
10778
10779
10780
10781
10782
10783
10784
10785
10786
10787
10788
10789
10790
10791
10792
10793
10794
10795
10796
10797
10798
10799
10800
10801
10802
10803
10804
10805
10806
10807
10808
10809
10810
10811
10812
10813
10814
10815
10816
10817
10818
10819
10820
10821
10822
10823
10824
10825
10826
10827
10828
10829
10830
10831
10832
10833
10834
10835
10836
10837
10838
10839
10840
10841
10842
10843
10844
10845
10846
10847
10848
10849
10850
10851
10852
10853
10854
10855
10856
10857
10858
10859
10860
10861
10862
10863
10864
10865
10866
10867
10868
10869
10870
10871
10872
10873
10874
10875
10876
10877
10878
10879
10880
10881
10882
10883
10884
10885
10886
10887
10888
10889
10890
10891
10892
10893
10894
10895
10896
10897
10898
10899
10900
10901
10902
10903
10904
10905
10906
10907
10908
10909
10910
10911
10912
10913
10914
10915
10916
10917
10918
10919
10920
10921
10922
10923
10924
10925
10926
10927
10928
10929
10930
10931
10932
10933
10934
10935
10936
10937
10938
10939
10940
10941
10942
10943
10944
10945
10946
10947
10948
10949
10950
10951
10952
10953
10954
10955
10956
10957
10958
10959
10960
10961
10962
10963
10964
10965
10966
10967
10968
10969
10970
10971
10972
10973
10974
10975
10976
10977
10978
10979
10980
10981
10982
10983
10984
10985
10986
10987
10988
10989
10990
10991
10992
10993
10994
10995
10996
10997
10998
10999
11000

FIGURE 4 (Sheet 2 of 3)

[illegible]

FIGURE 4
(sheet 3 of 3)

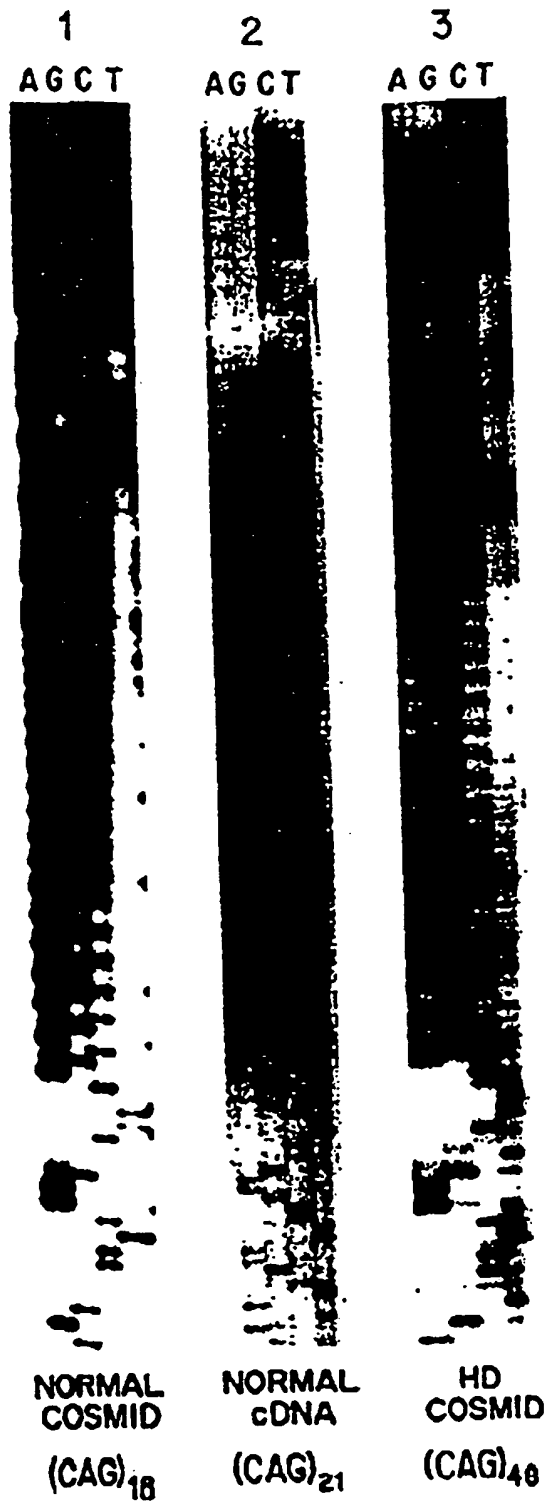


FIGURE 5

FIGURE 6

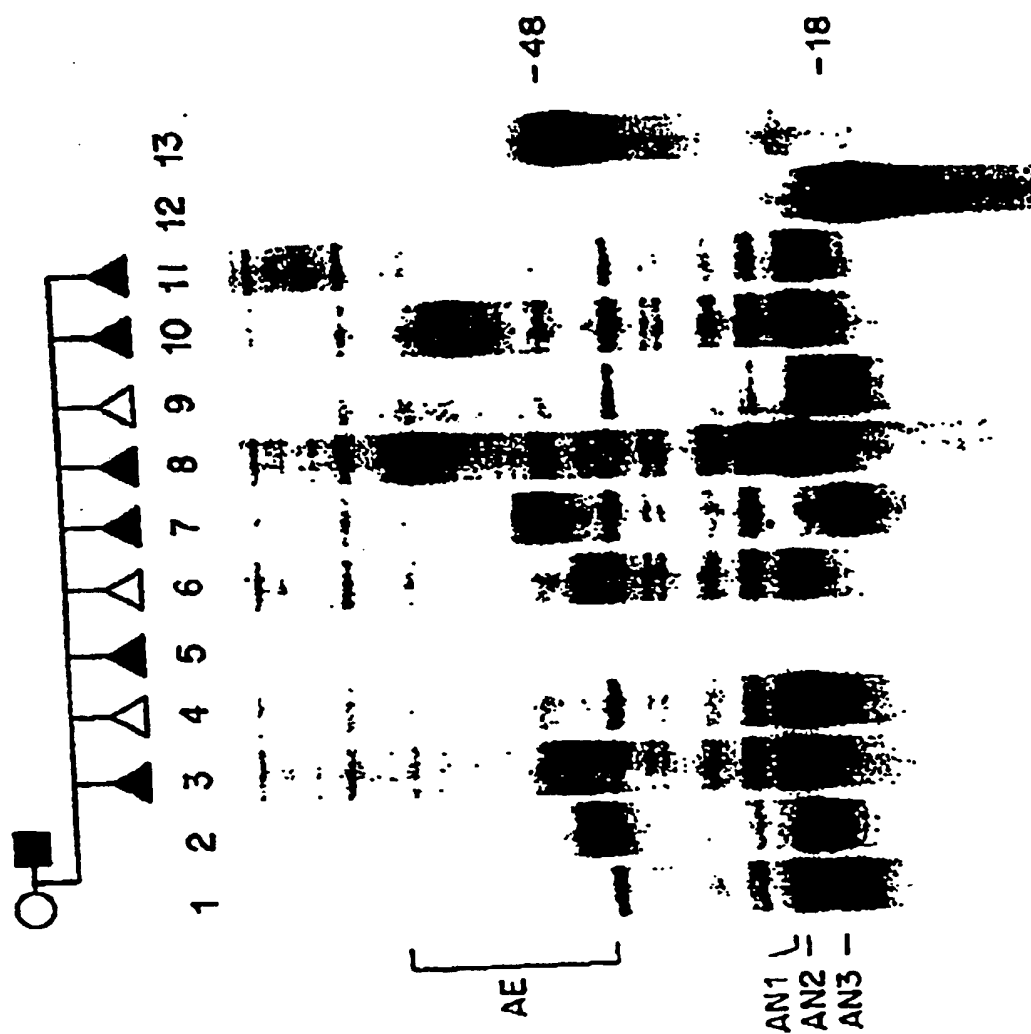
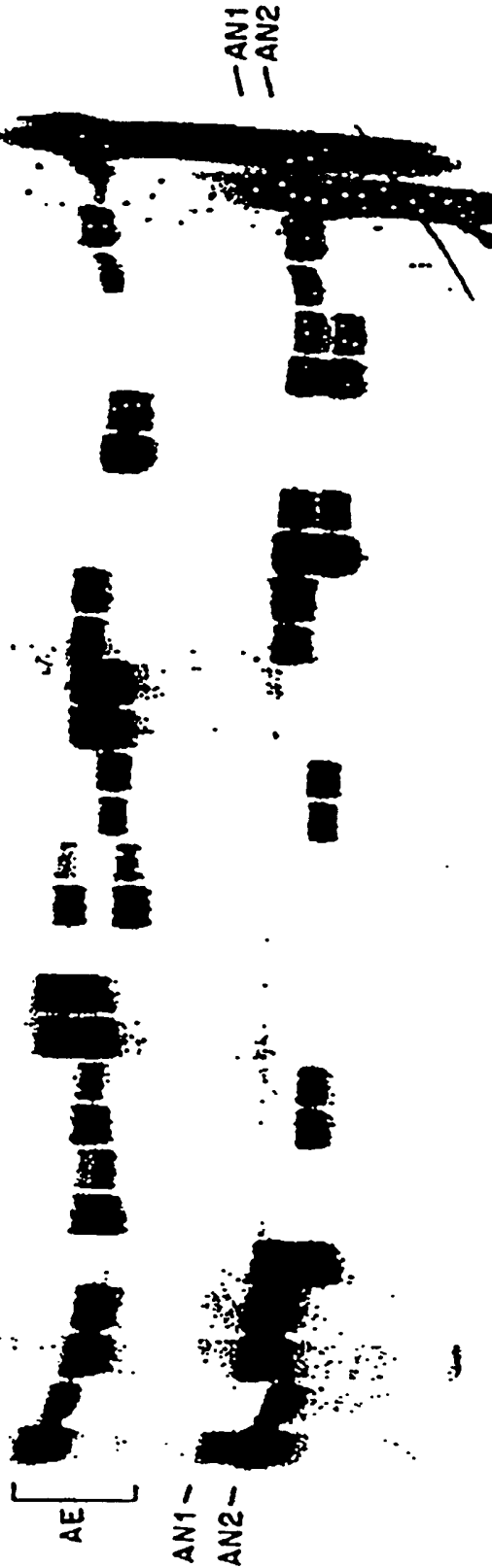
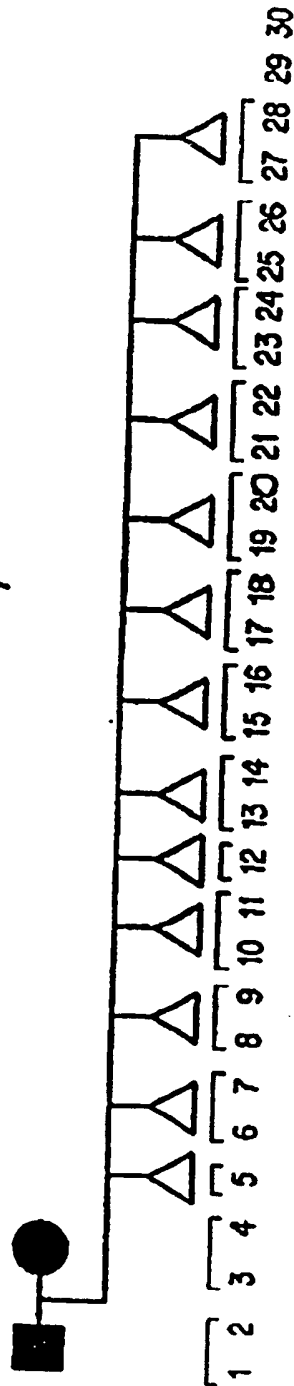


FIGURE 7



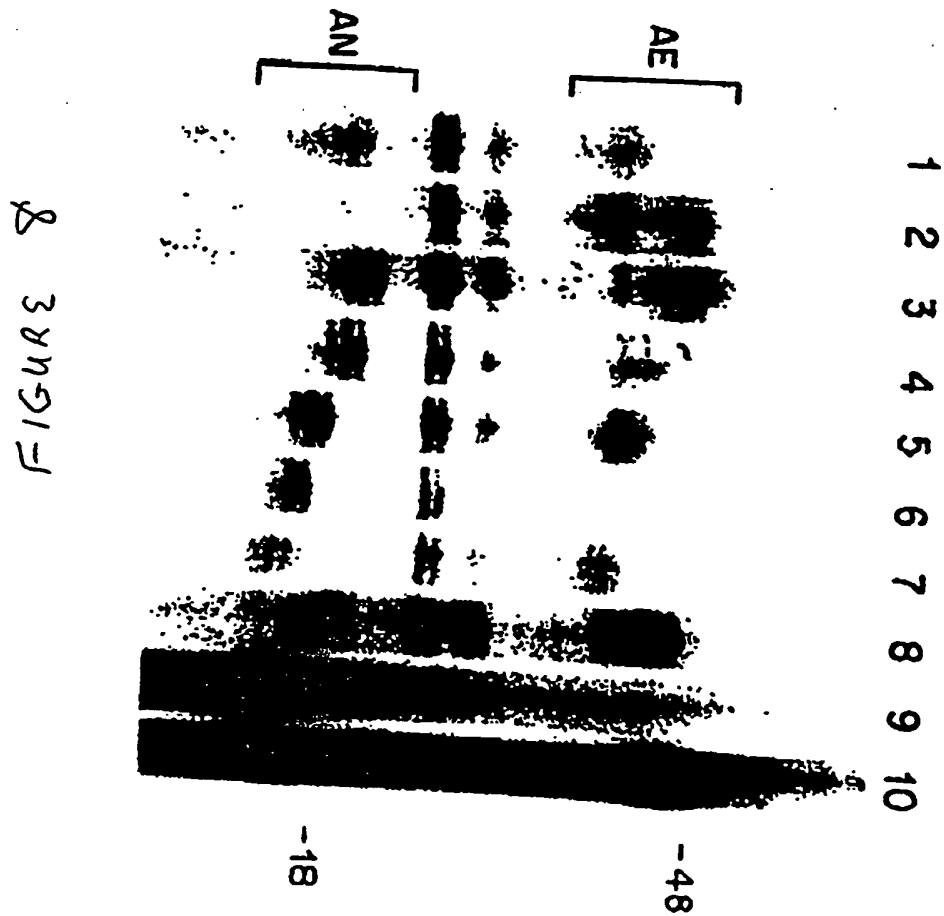


FIGURE 9

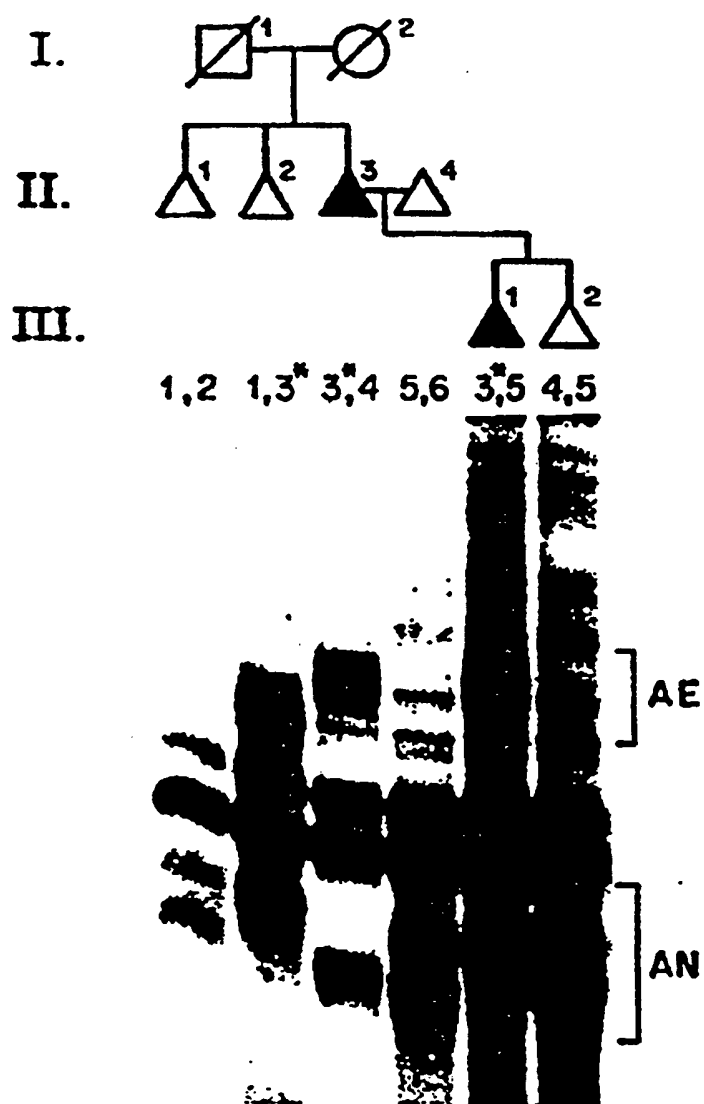


FIGURE 10

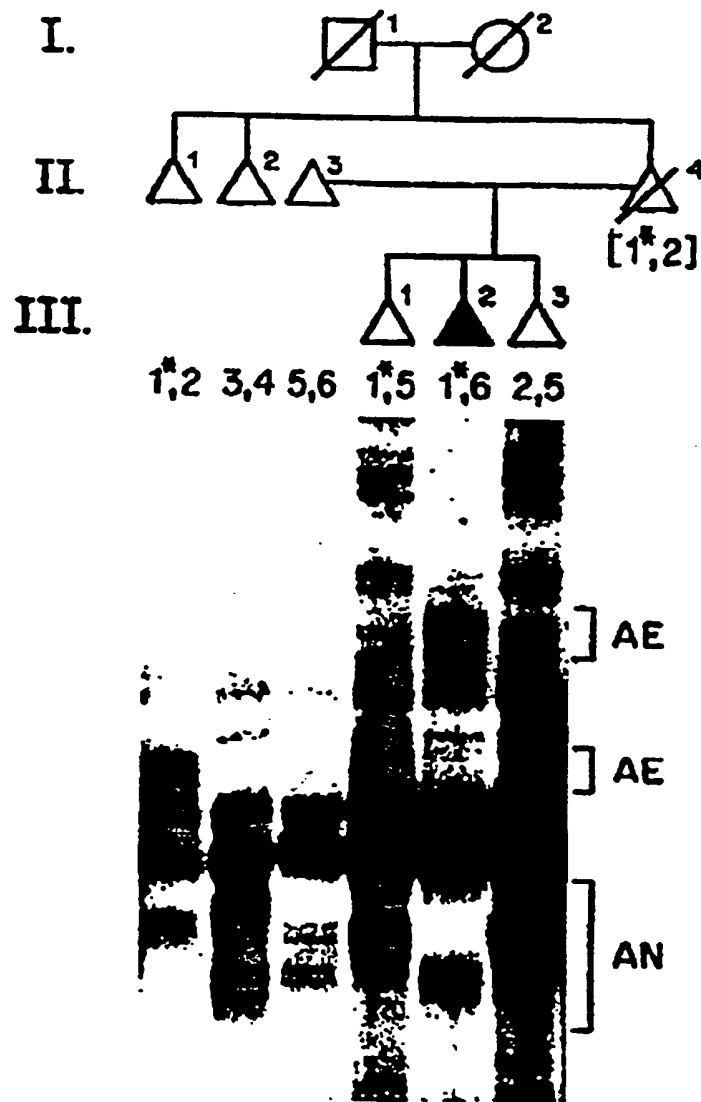


FIGURE 11

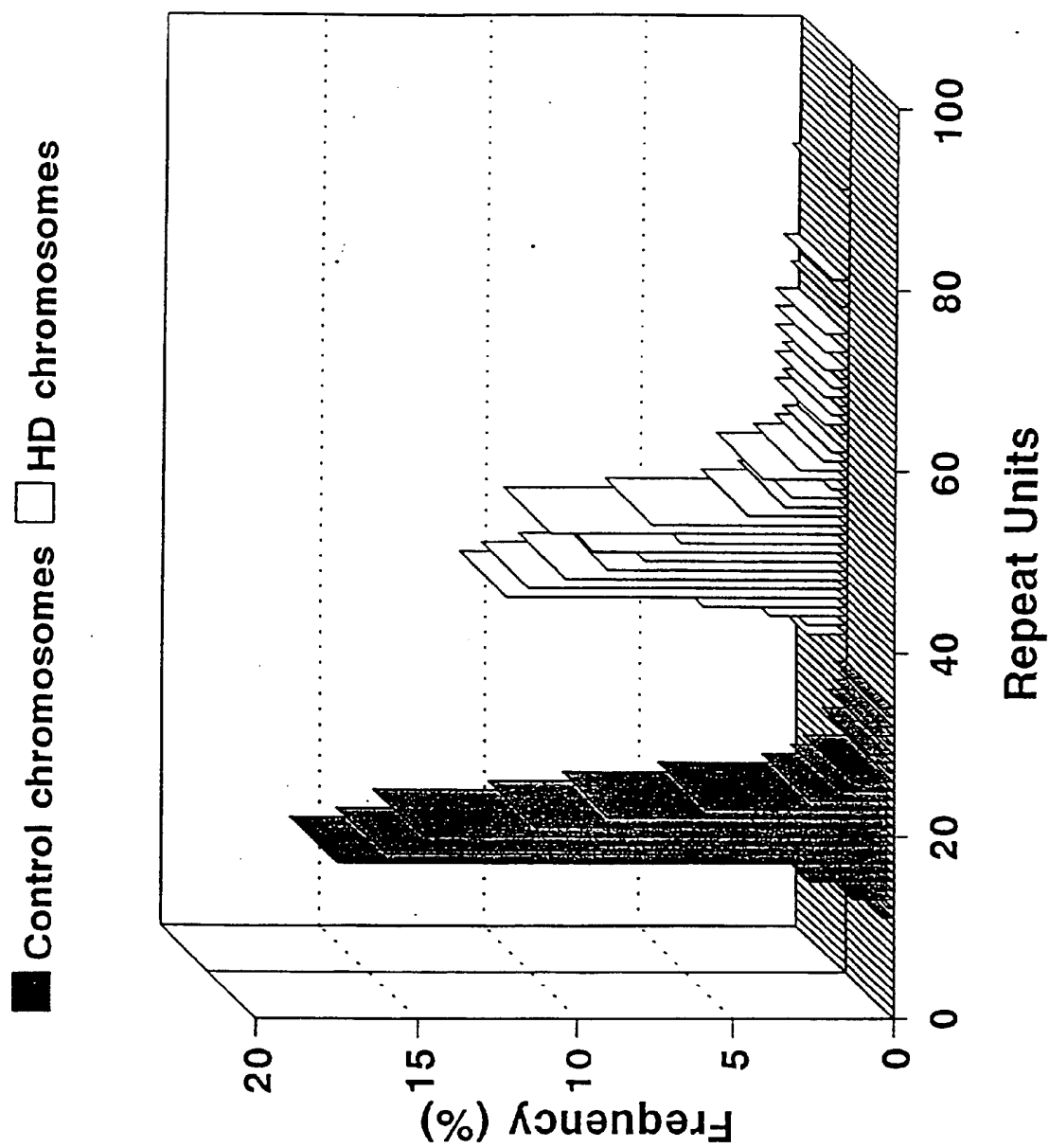


FIGURE 12 (PANEL A)

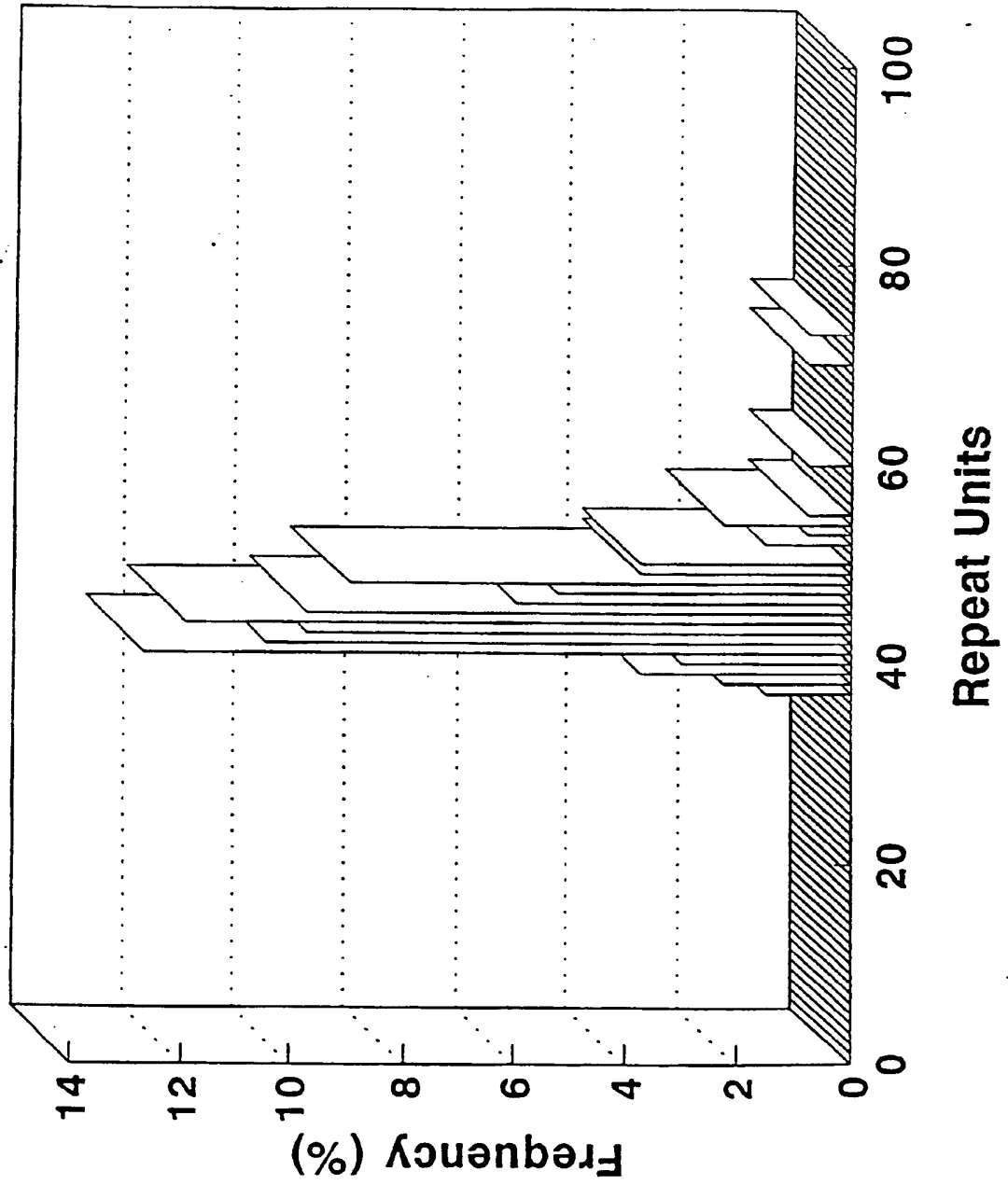


FIGURE 12 cont. (PANEL B)

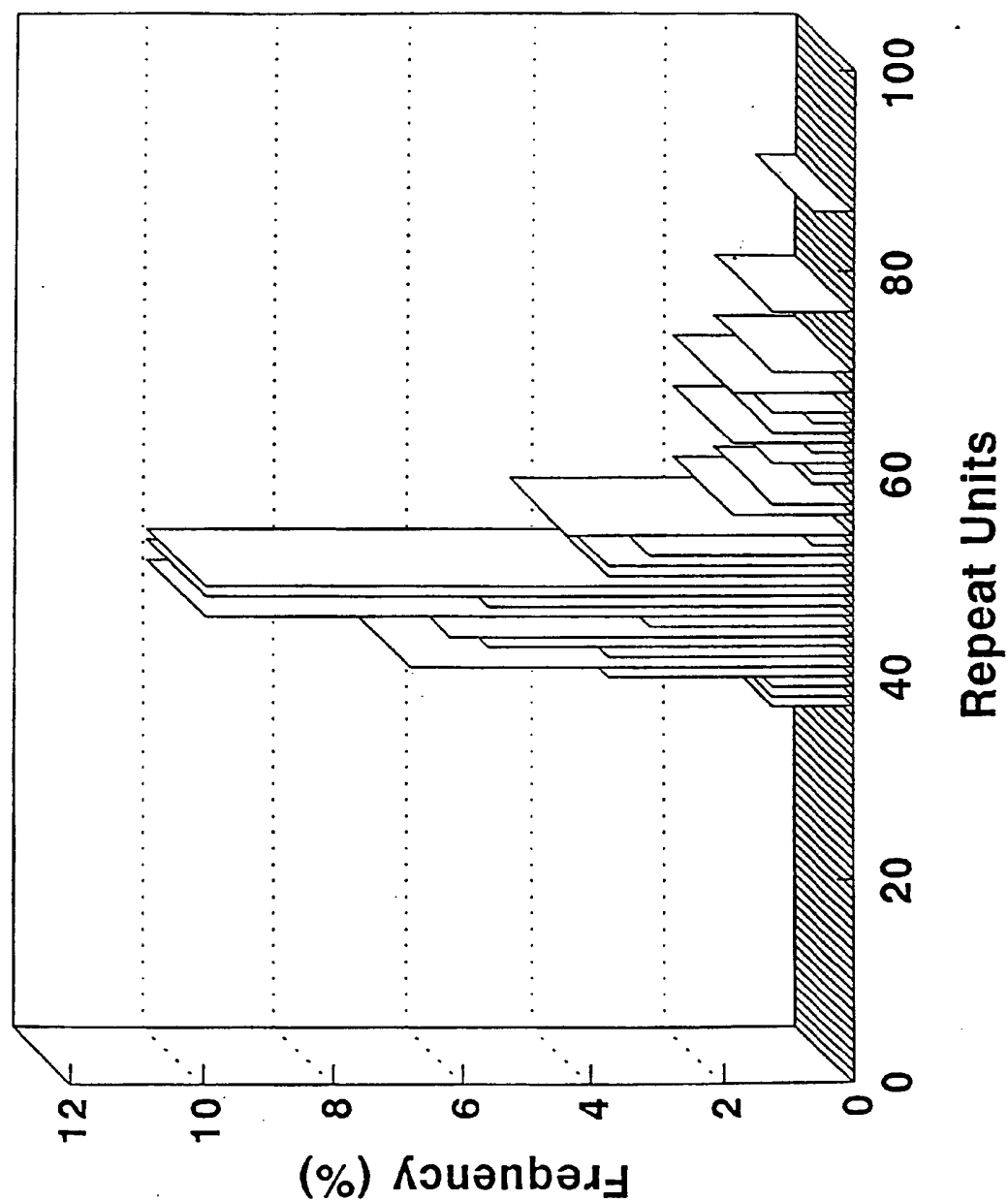


FIGURE 13 (PANEL A)

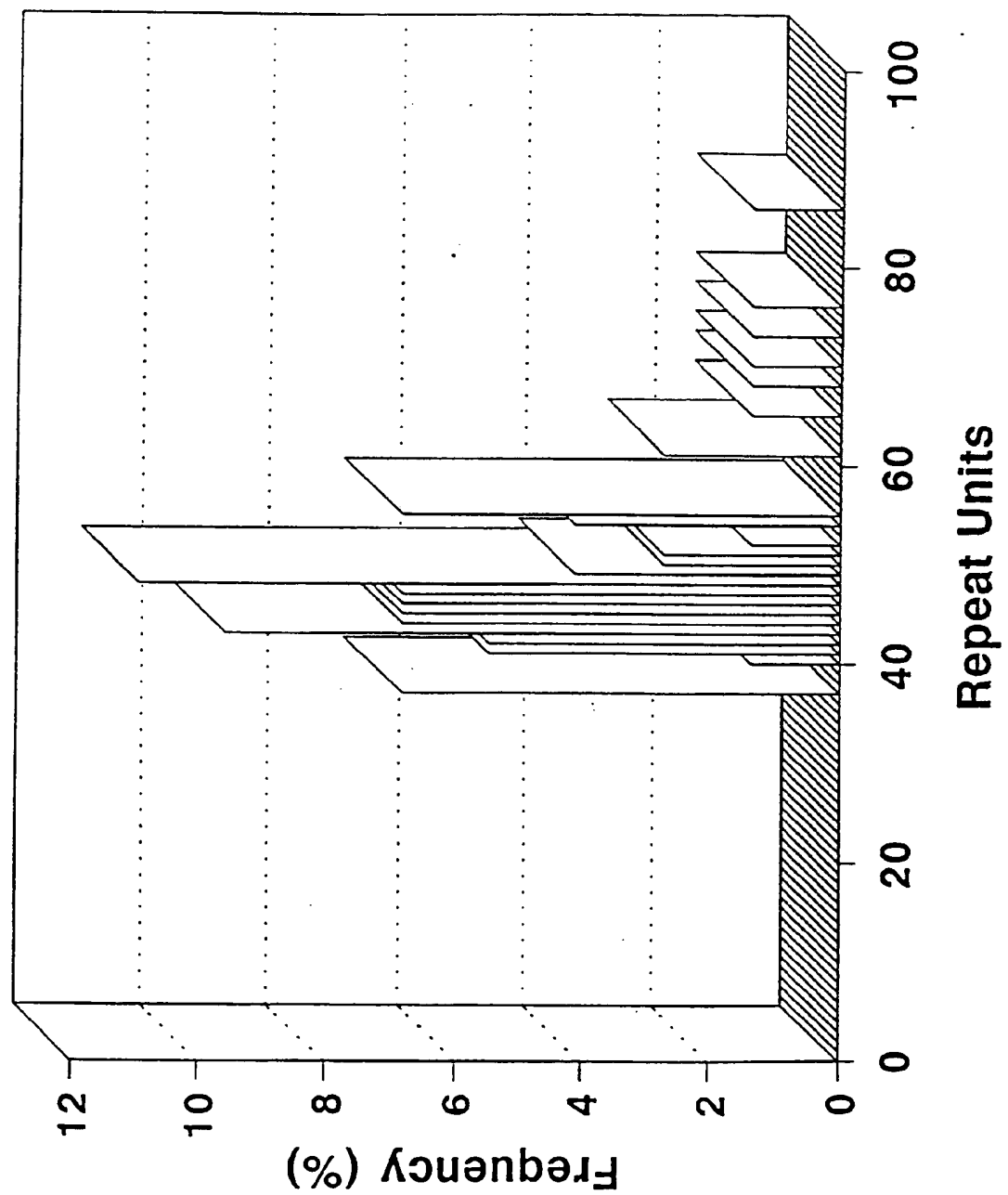


FIGURE 13 cont. (PANEL B)

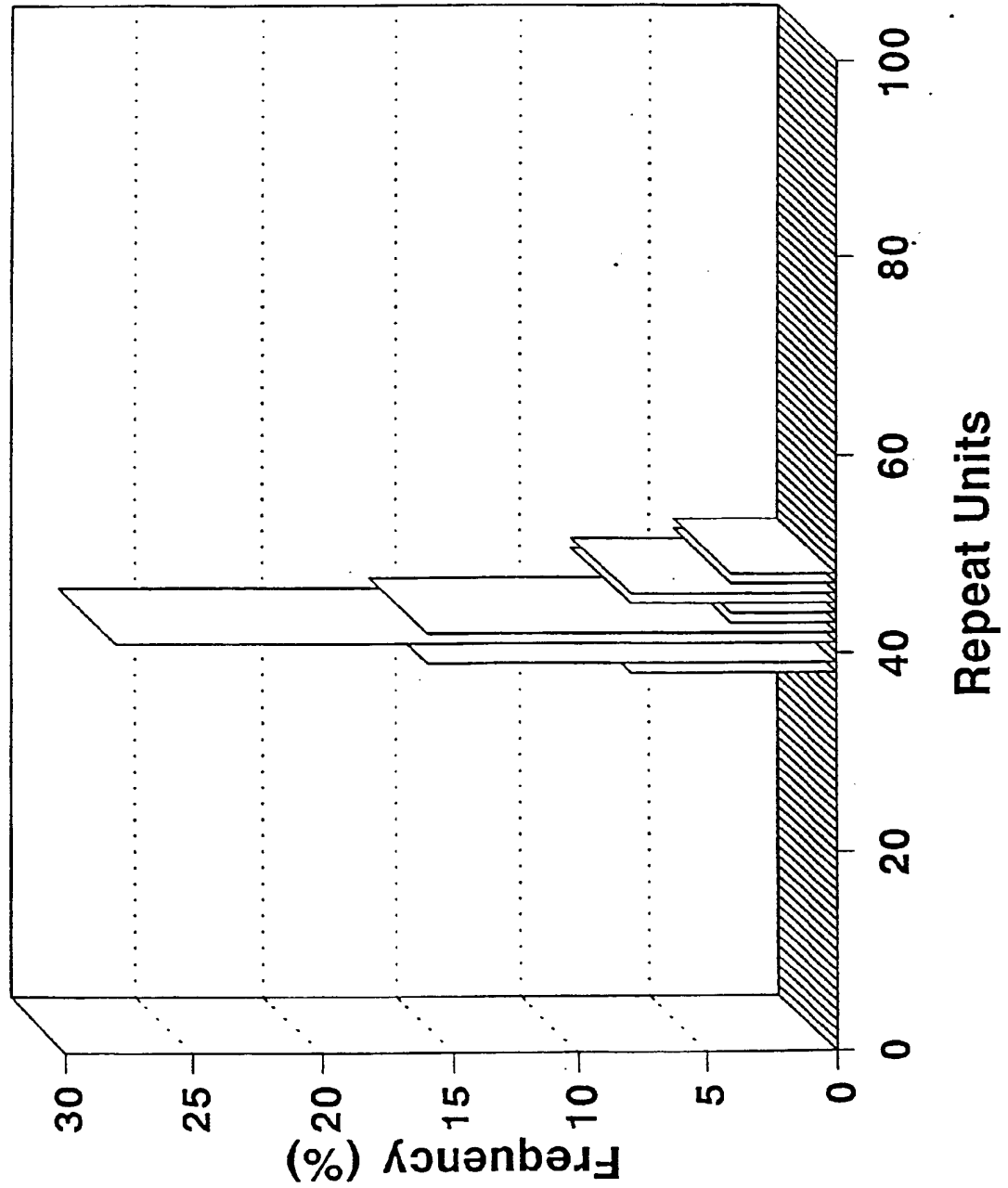


FIGURE 13 cont. (PANEL C)

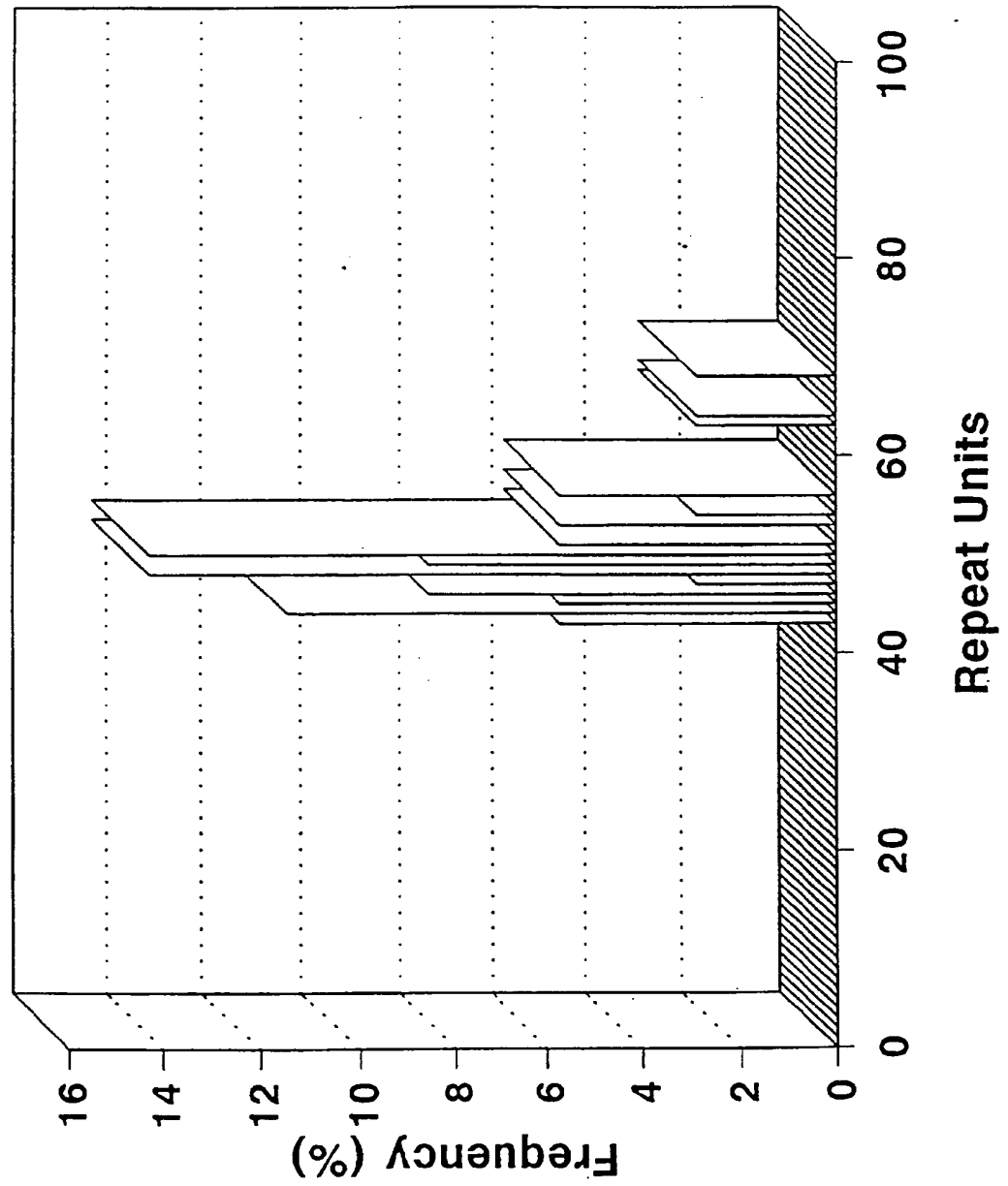


FIGURE 14 (PANEL A)

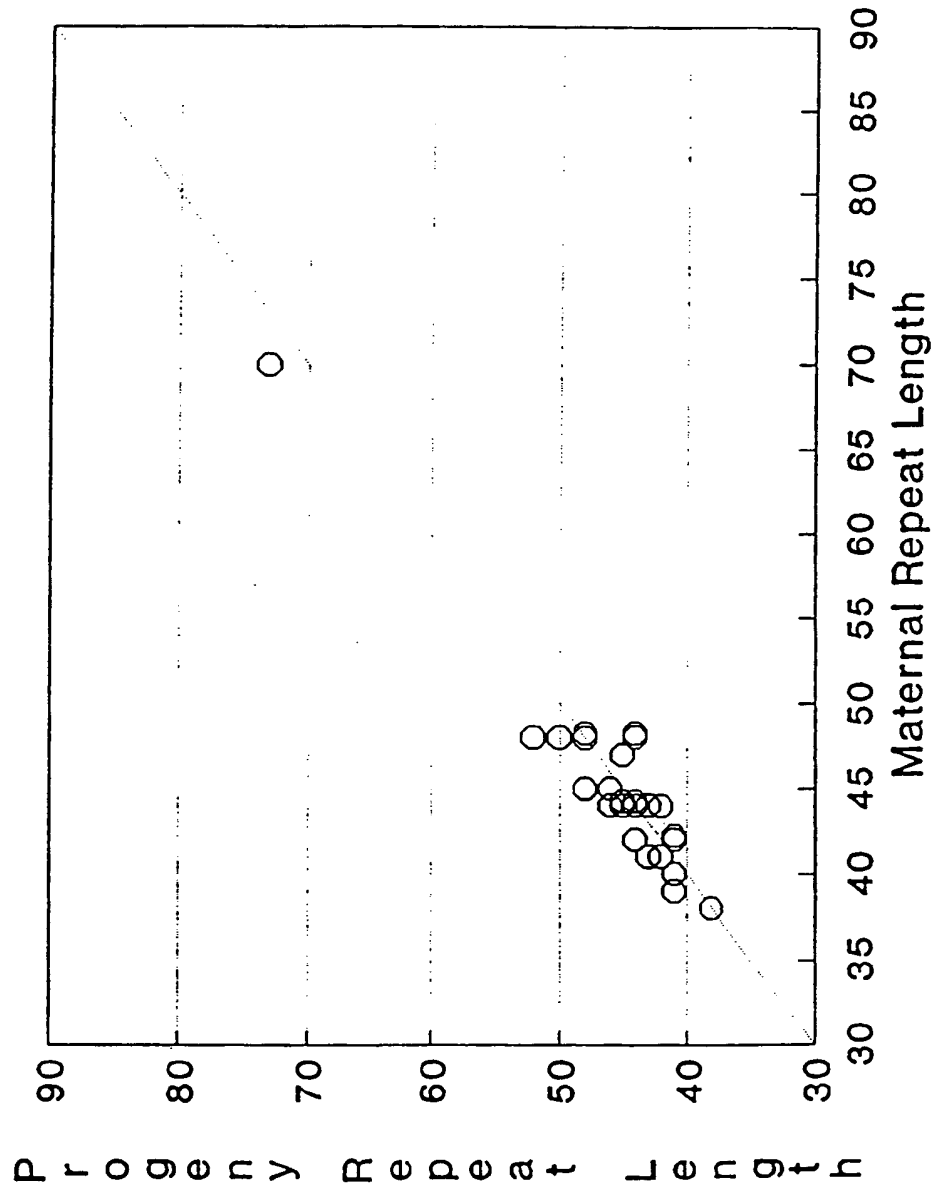


FIGURE 14 cont. (PANEL B)

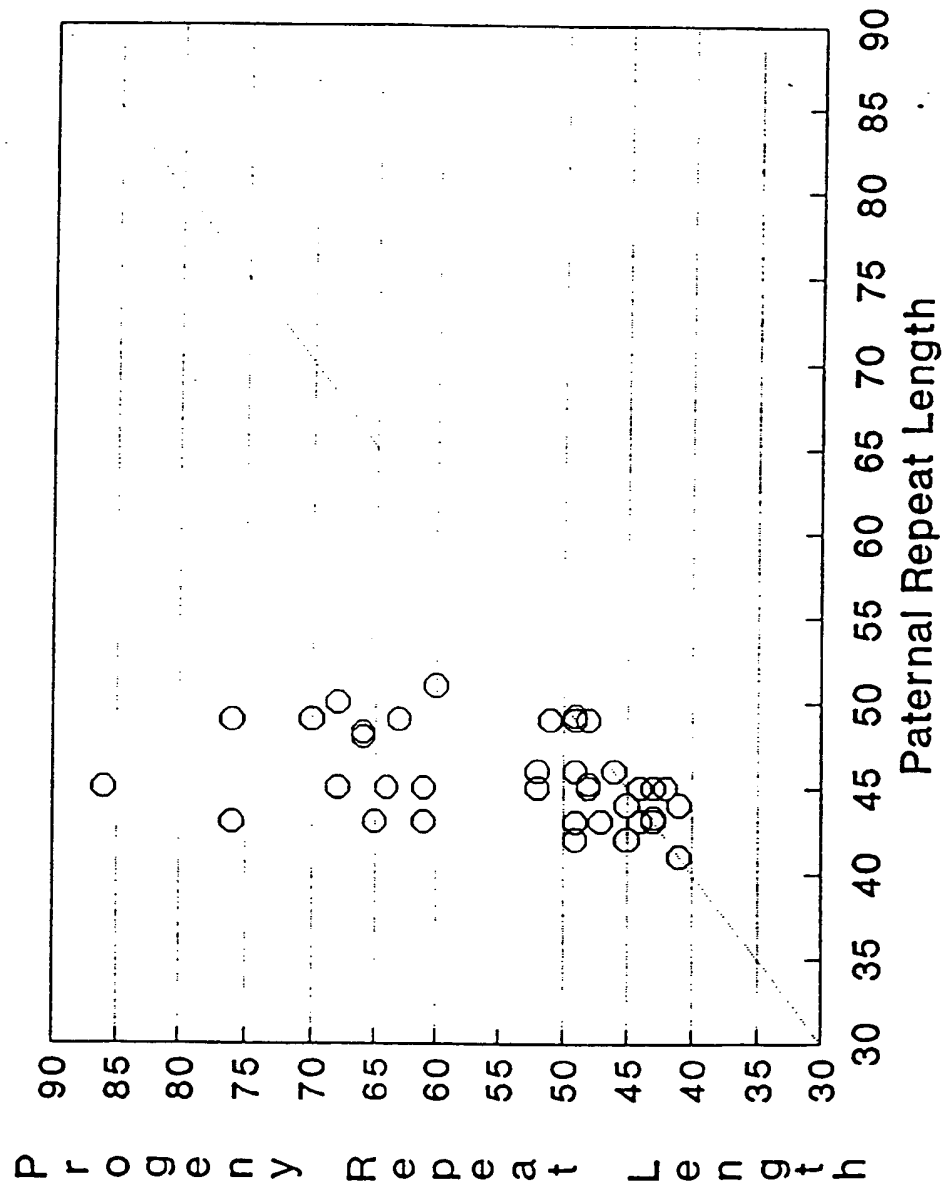


FIGURE 15

1	2	3	4	5
┌───┐	┌───┐	┌───┐	┌───┐	┌───┐
S L	S L	S L	S L	S L

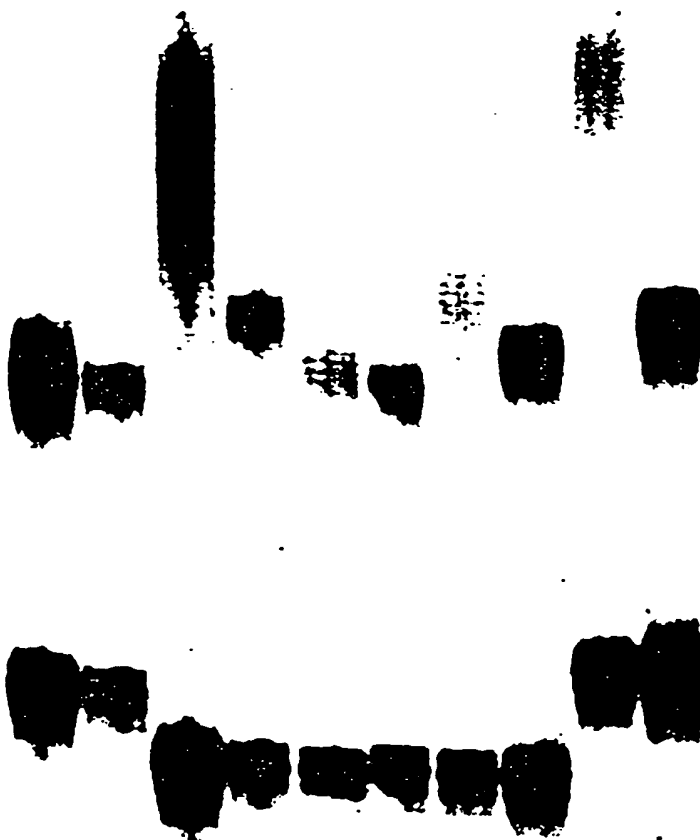
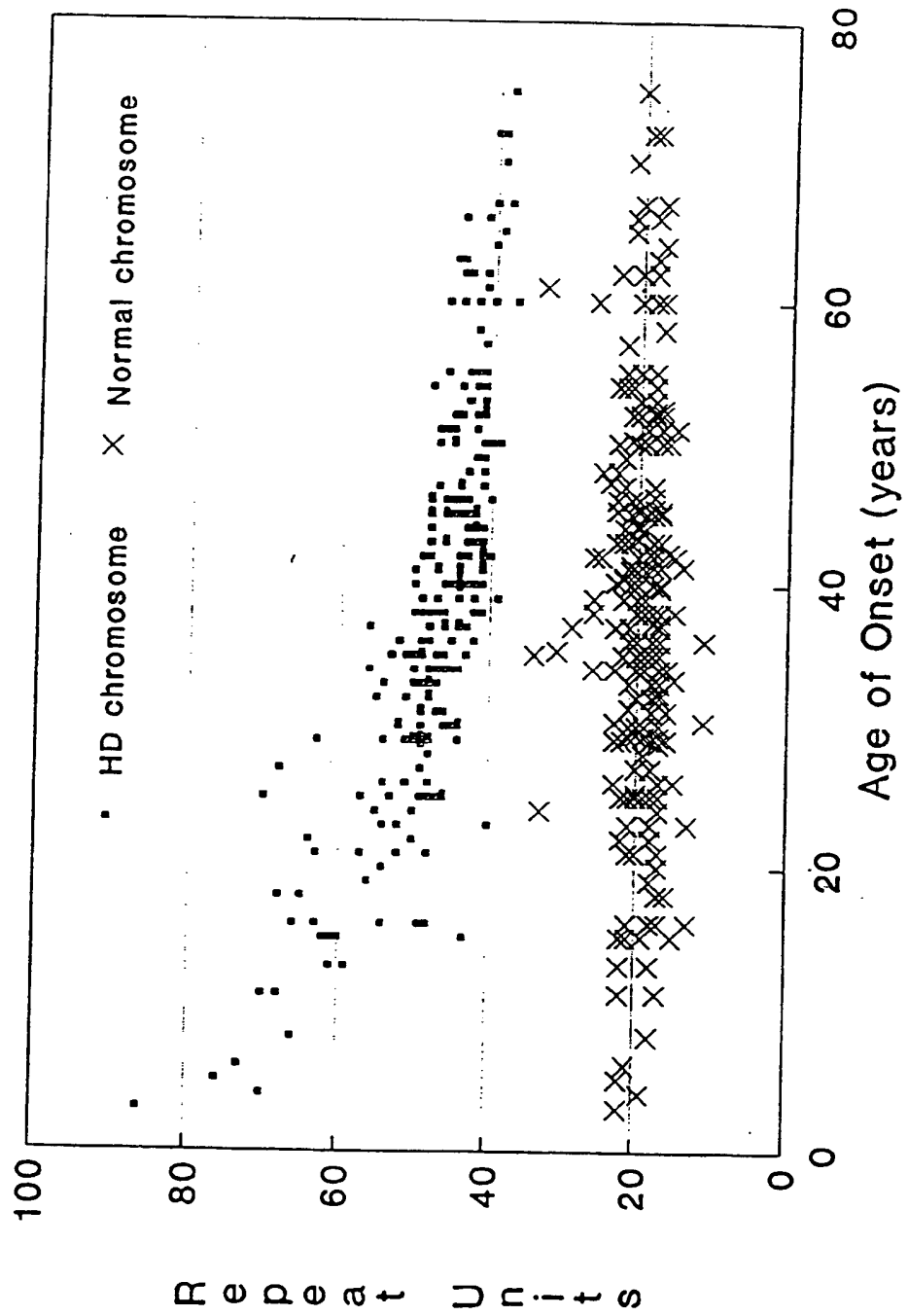


FIGURE 16



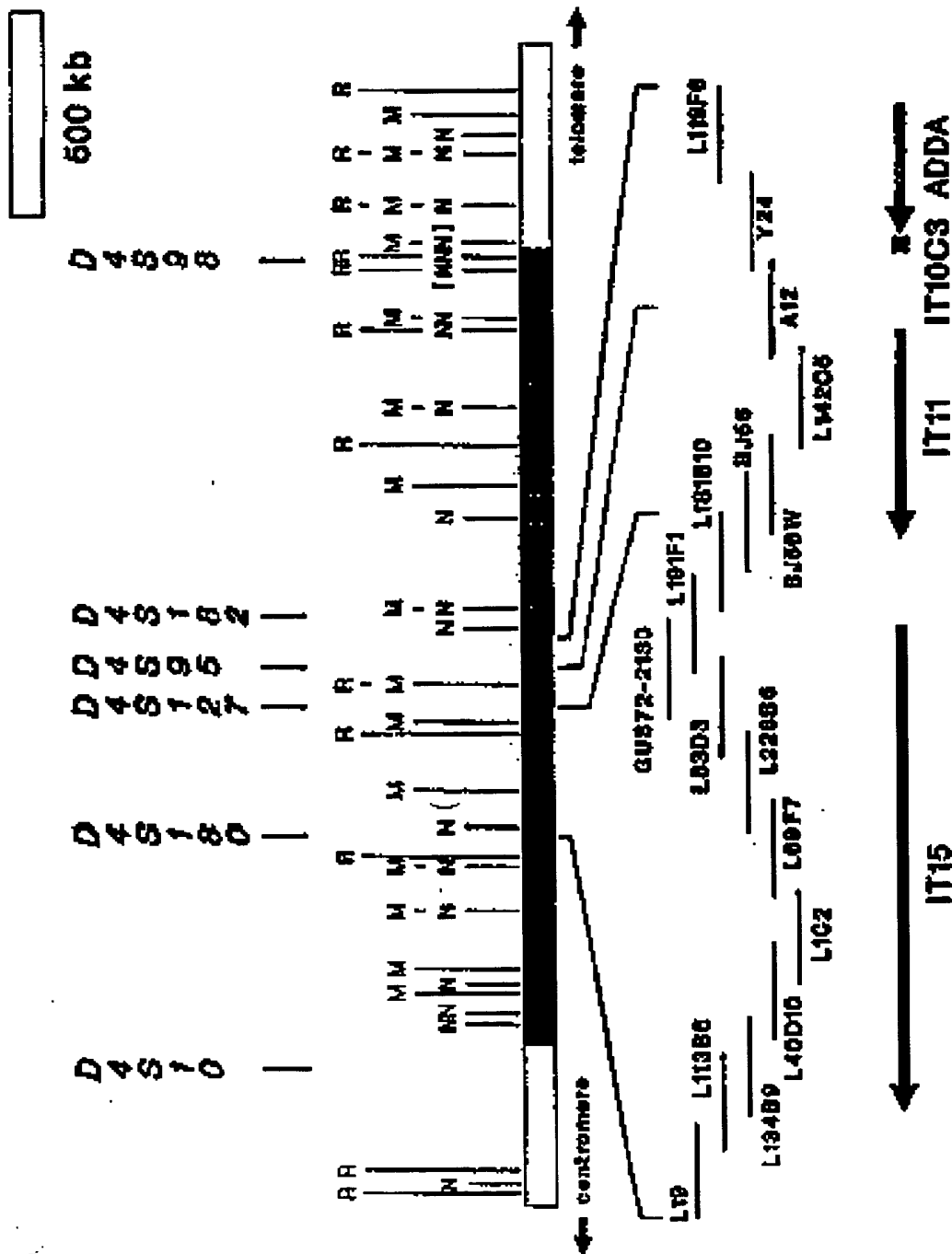


FIGURE 1

1 2 3



— 28S



— 18S

Figure 2

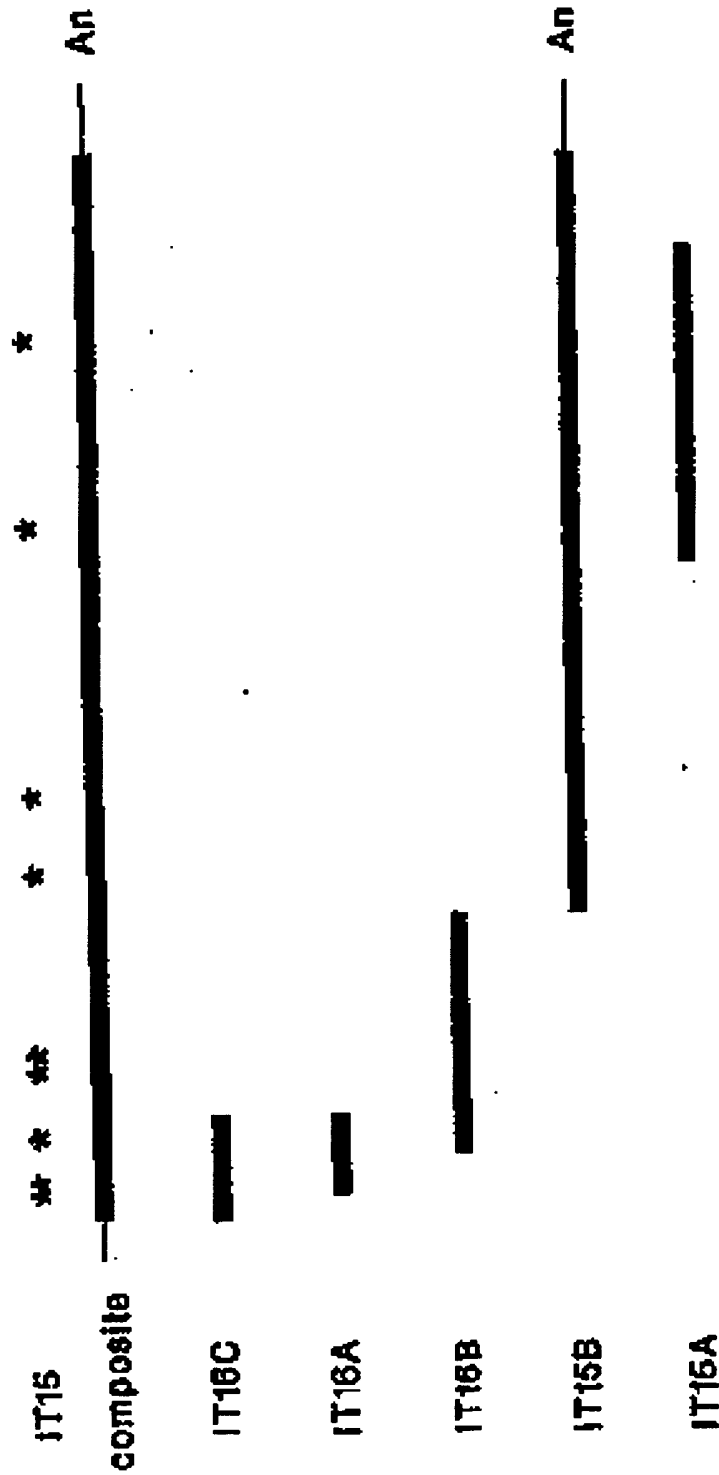


Figure 3

[illegible]

FIGURE 4 (Sheet 1 of 3)

1050
1051
1052
1053
1054
1055
1056
1057
1058
1059
1060
1061
1062
1063
1064
1065
1066
1067
1068
1069
1070
1071
1072
1073
1074
1075
1076
1077
1078
1079
1080
1081
1082
1083
1084
1085
1086
1087
1088
1089
1090
1091
1092
1093
1094
1095
1096
1097
1098
1099
1100
1101
1102
1103
1104
1105
1106
1107
1108
1109
1110
1111
1112
1113
1114
1115
1116
1117
1118
1119
1120
1121
1122
1123
1124
1125
1126
1127
1128
1129
1130
1131
1132
1133
1134
1135
1136
1137
1138
1139
1140
1141
1142
1143
1144
1145
1146
1147
1148
1149
1150
1151
1152
1153
1154
1155
1156
1157
1158
1159
1160
1161
1162
1163
1164
1165
1166
1167
1168
1169
1170
1171
1172
1173
1174
1175
1176
1177
1178
1179
1180
1181
1182
1183
1184
1185
1186
1187
1188
1189
1190
1191
1192
1193
1194
1195
1196
1197
1198
1199
1200
1201
1202
1203
1204
1205
1206
1207
1208
1209
1210
1211
1212
1213
1214
1215
1216
1217
1218
1219
1220
1221
1222
1223
1224
1225
1226
1227
1228
1229
1230
1231
1232
1233
1234
1235
1236
1237
1238
1239
1240
1241
1242
1243
1244
1245
1246
1247
1248
1249
1250
1251
1252
1253
1254
1255
1256
1257
1258
1259
1260
1261
1262
1263
1264
1265
1266
1267
1268
1269
1270
1271
1272
1273
1274
1275
1276
1277
1278
1279
1280
1281
1282
1283
1284
1285
1286
1287
1288
1289
1290
1291
1292
1293
1294
1295
1296
1297
1298
1299
1300
1301
1302
1303
1304
1305
1306
1307
1308
1309
1310
1311
1312
1313
1314
1315
1316
1317
1318
1319
1320
1321
1322
1323
1324
1325
1326
1327
1328
1329
1330
1331
1332
1333
1334
1335
1336
1337
1338
1339
1340
1341
1342
1343
1344
1345
1346
1347
1348
1349
1350
1351
1352
1353
1354
1355
1356
1357
1358
1359
1360
1361
1362
1363
1364
1365
1366
1367
1368
1369
1370
1371
1372
1373
1374
1375
1376
1377
1378
1379
1380
1381
1382
1383
1384
1385
1386
1387
1388
1389
1390
1391
1392
1393
1394
1395
1396
1397
1398
1399
1400
1401
1402
1403
1404
1405
1406
1407
1408
1409
1410
1411
1412
1413
1414
1415
1416
1417
1418
1419
1420
1421
1422
1423
1424
1425
1426
1427
1428
1429
1430
1431
1432
1433
1434
1435
1436
1437
1438
1439
1440
1441
1442
1443
1444
1445
1446
1447
1448
1449
1450
1451
1452
1453
1454
1455
1456
1457
1458
1459
1460
1461
1462
1463
1464
1465
1466
1467
1468
1469
1470
1471
1472
1473
1474
1475
1476
1477
1478
1479
1480
1481
1482
1483
1484
1485
1486
1487
1488
1489
1490
1491
1492
1493
1494
1495
1496
1497
1498
1499
1500
1501
1502
1503
1504
1505
1506
1507
1508
1509
1510
1511
1512
1513
1514
1515
1516
1517
1518
1519
1520
1521
1522
1523
1524
1525
1526
1527
1528
1529
1530
1531
1532
1533
1534
1535
1536
1537
1538
1539
1540
1541
1542
1543
1544
1545
1546
1547
1548
1549
1550
1551
1552
1553
1554
1555
1556
1557
1558
1559
1560
1561
1562
1563
1564
1565
1566
1567
1568
1569
1570
1571
1572
1573
1574
1575
1576
1577
1578
1579
1580
1581
1582
1583
1584
1585
1586
1587
1588
1589
1590
1591
1592
1593
1594
1595
1596
1597
1598
1599
1600
1601
1602
1603
1604
1605
1606
1607
1608
1609
1610
1611
1612
1613
1614
1615
1616
1617
1618
1619
1620
1621
1622
1623
1624
1625
1626
1627
1628
1629
1630
1631
1632
1633
1634
1635
1636
1637
1638
1639
1640
1641
1642
1643
1644
1645
1646
1647
1648
1649
1650
1651
1652
1653
1654
1655
1656
1657
1658
1659
1660
1661
1662
1663
1664
1665
1666
1667
1668
1669
1670
1671
1672
1673
1674
1675
1676
1677
1678
1679
1680
1681
1682
1683
1684
1685
1686
1687
1688
1689
1690
1691
1692
1693
1694
1695
1696
1697
1698
1699
1700
1701
1702
1703
1704
1705
1706
1707
1708
1709
1710
1711
1712
1713
1714
1715
1716
1717
1718
1719
1720
1721
1722
1723
1724
1725
1726
1727
1728
1729
1730
1731
17

FIGURE 4 (Sheet 2 of 3)

[illegible]

FIGURE 4
(Sheet 3 of 3)



FIGURE 5

FIGURE 6

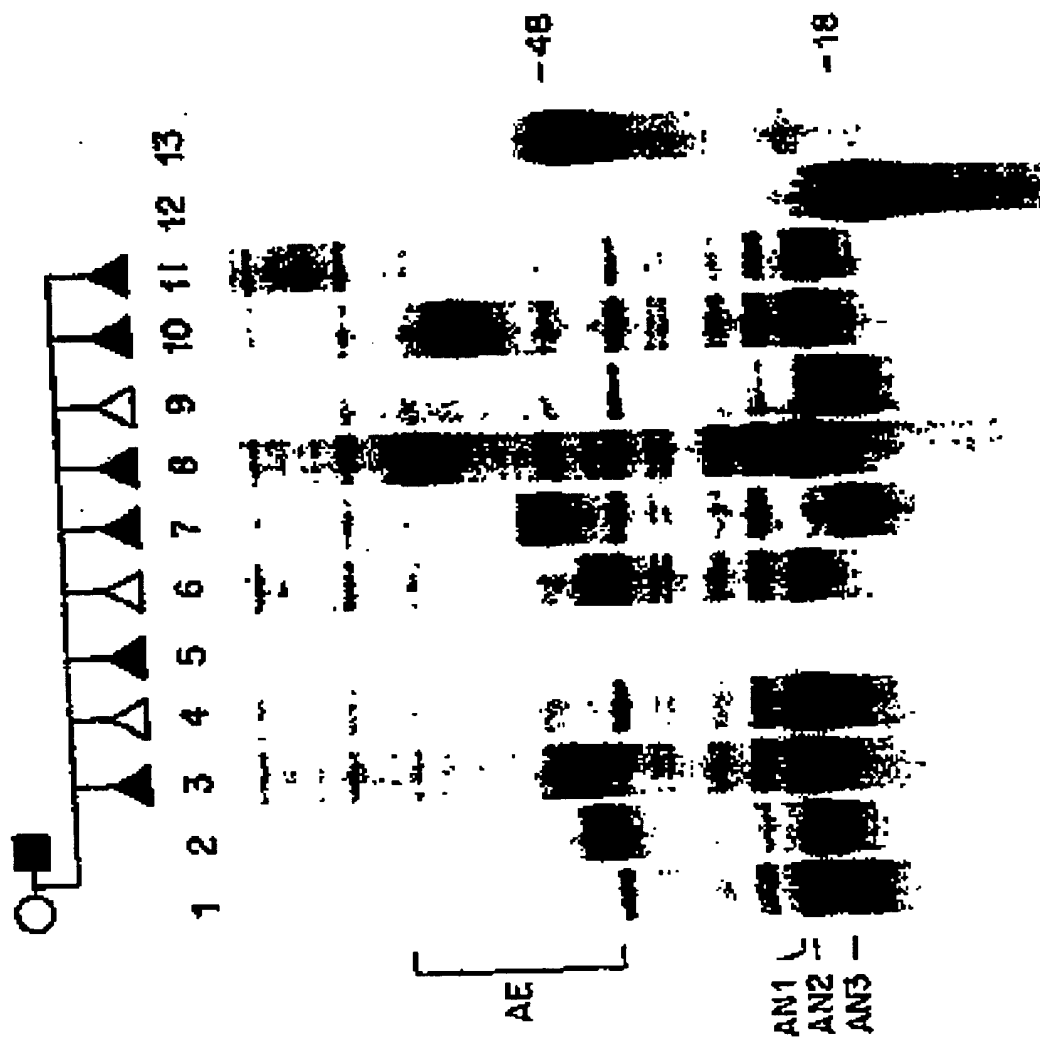
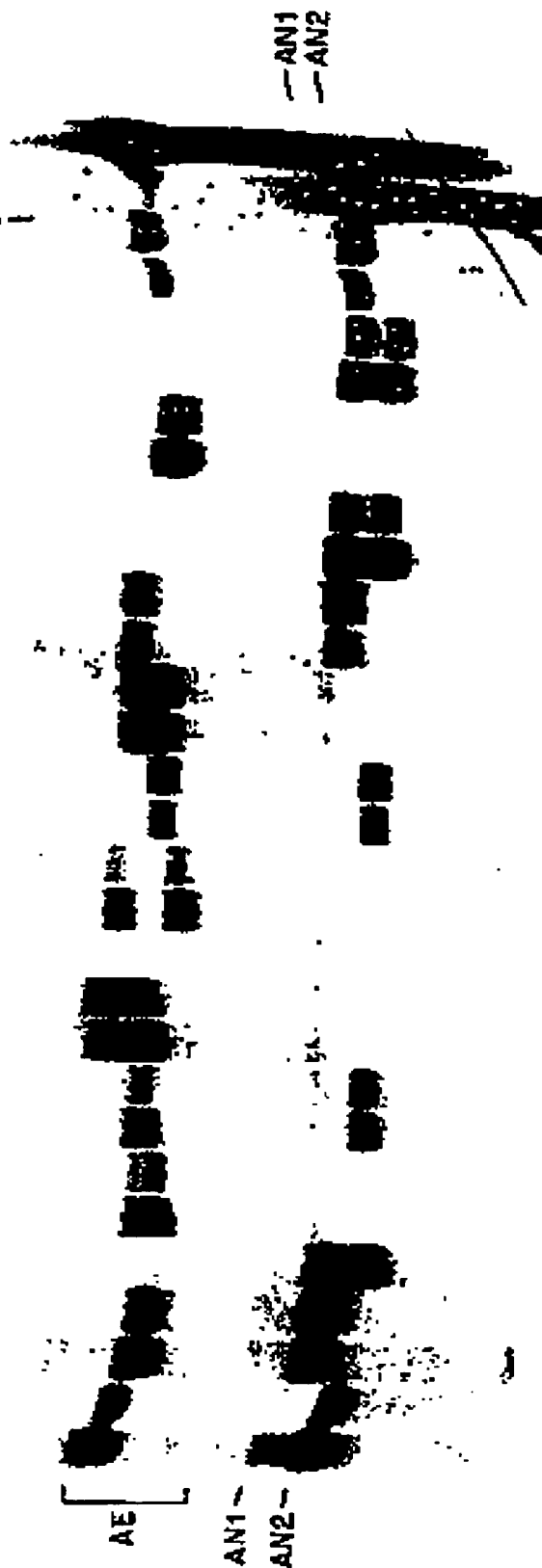
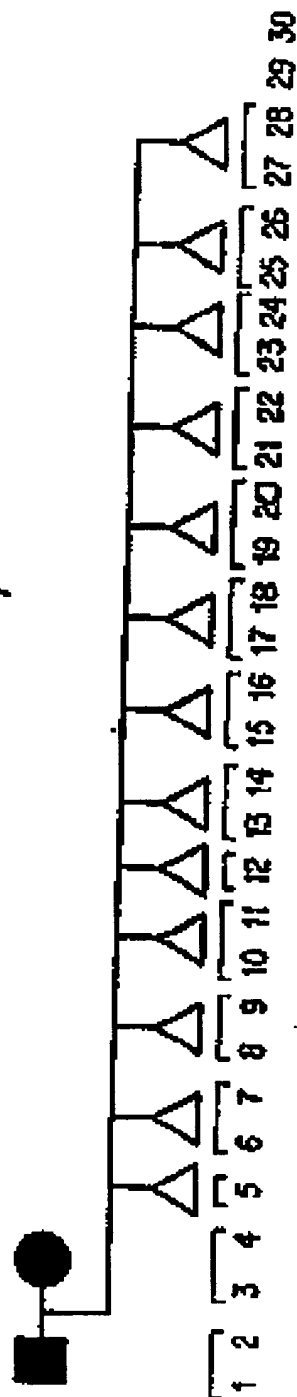


FIGURE 7



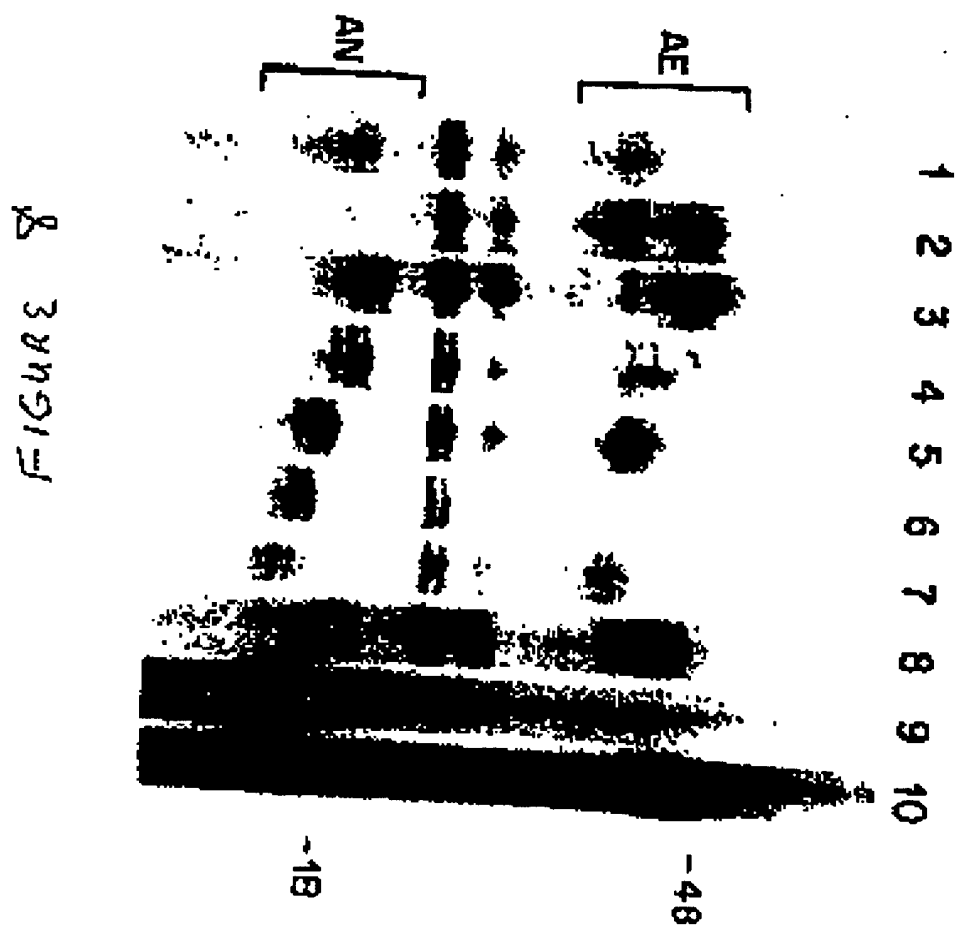


FIGURE 9

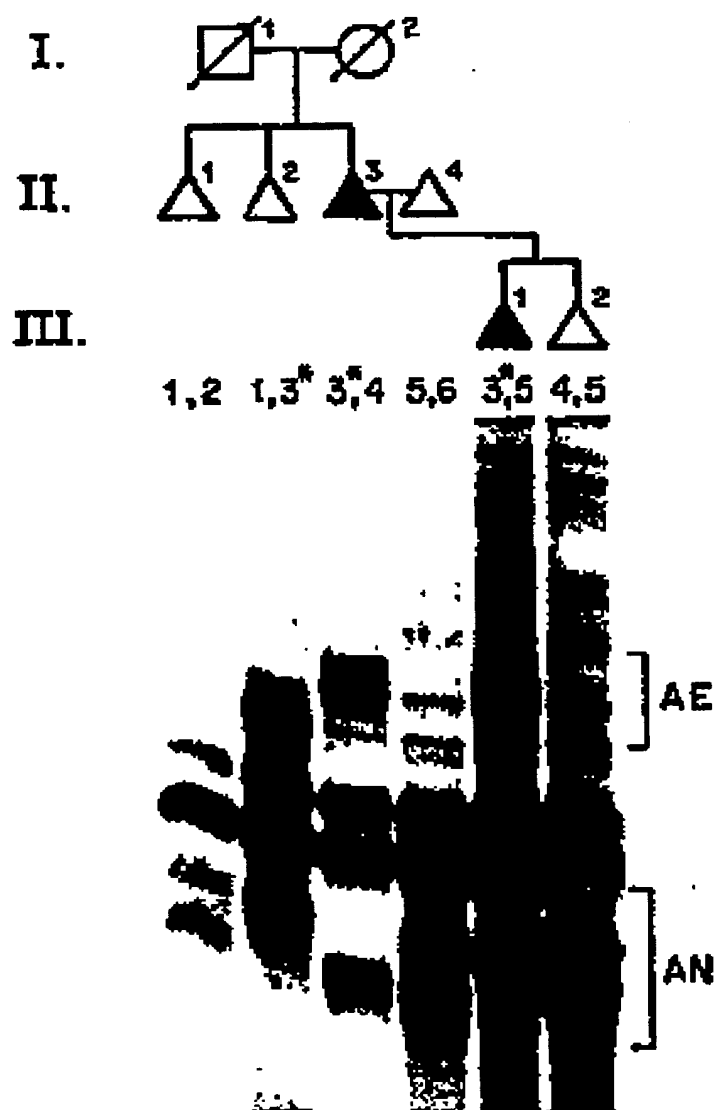


FIGURE 10

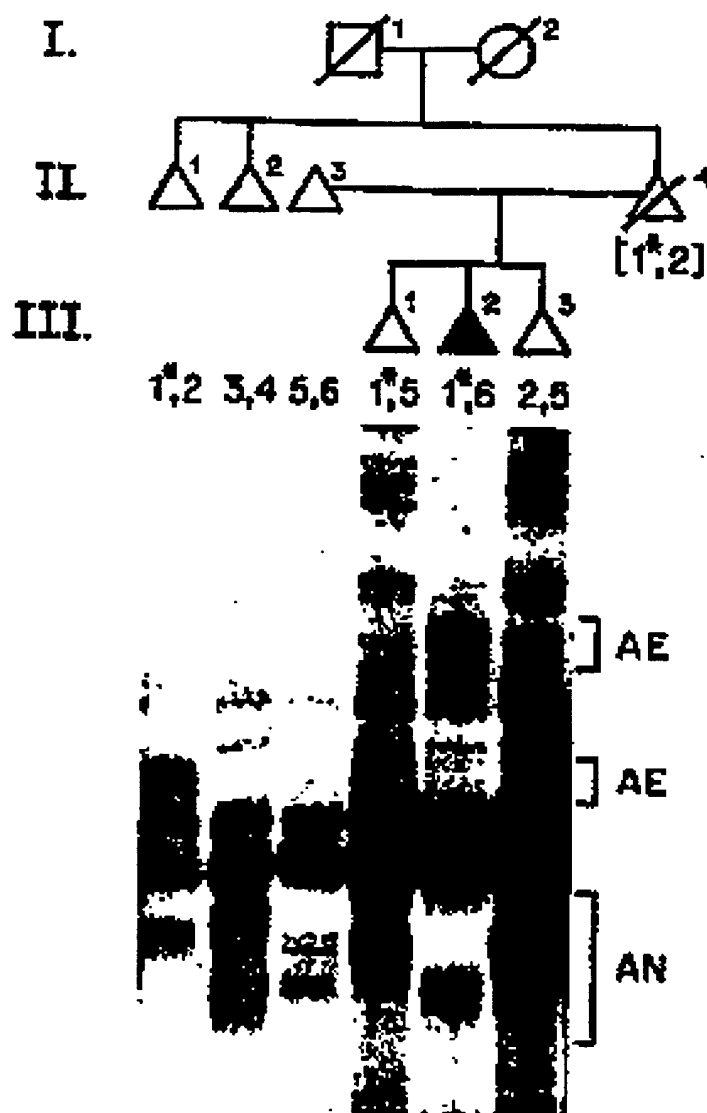


FIGURE 11

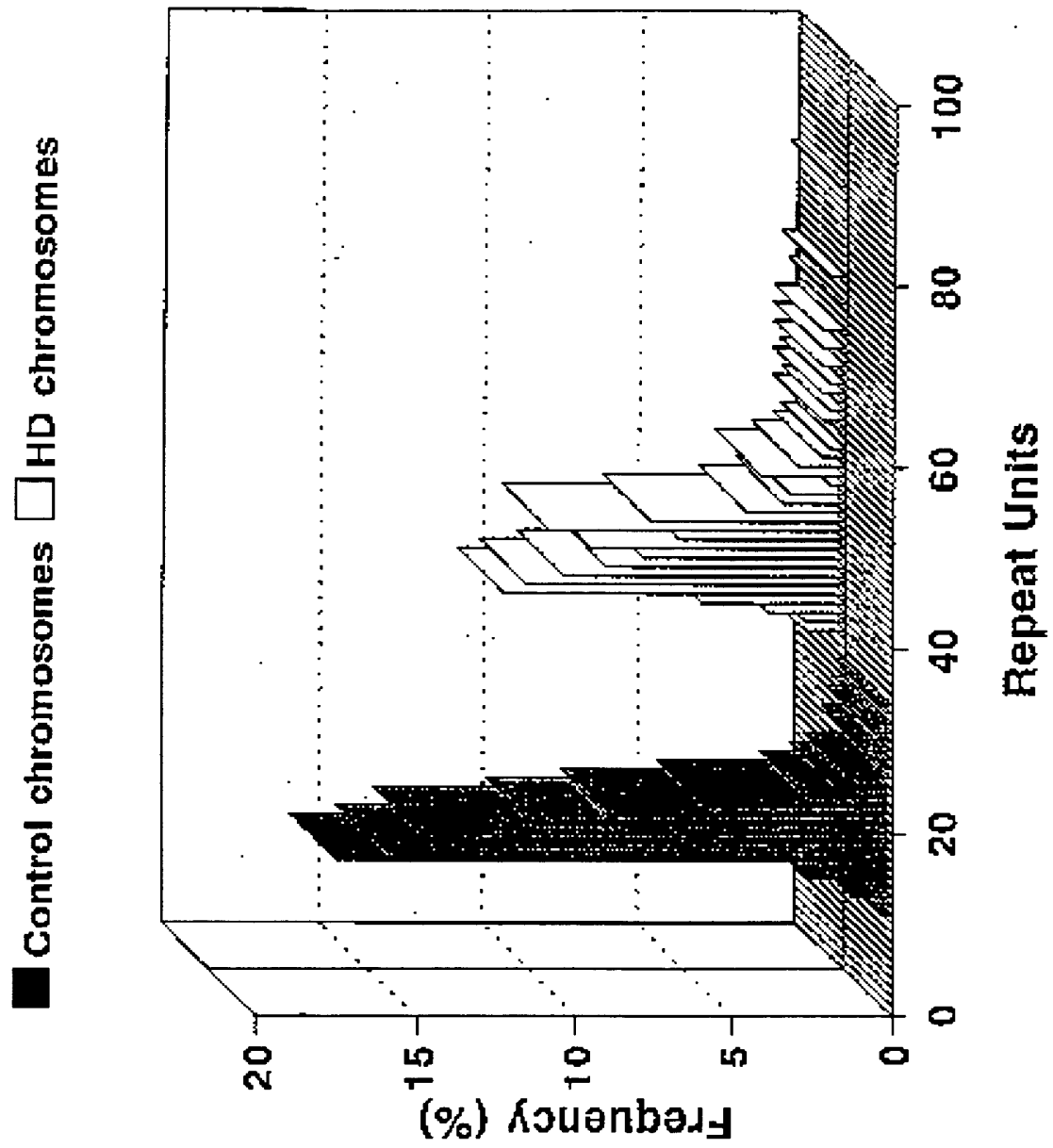


FIGURE 12 (PANEL A)

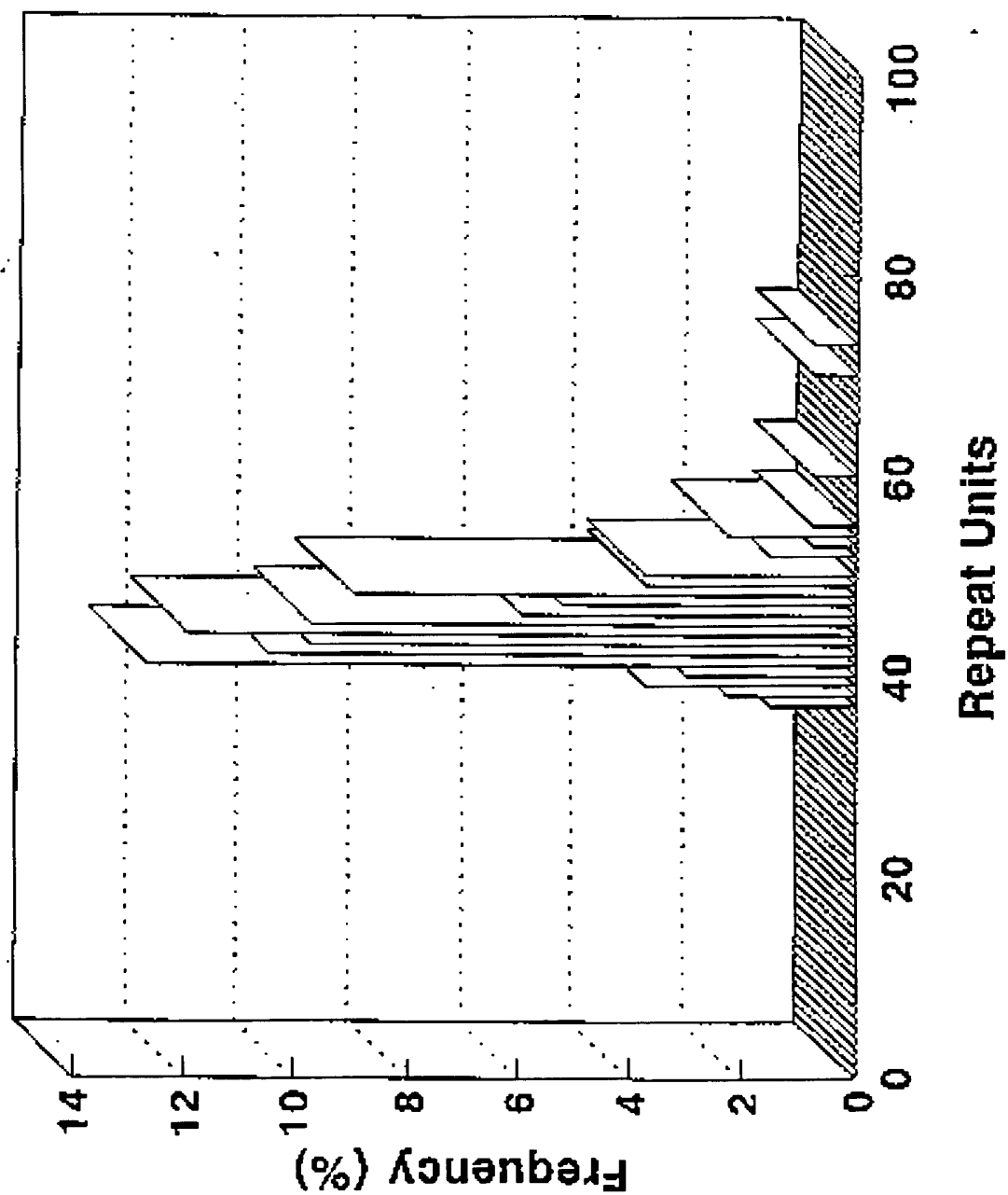


FIGURE 12 cont. (PANEL B)

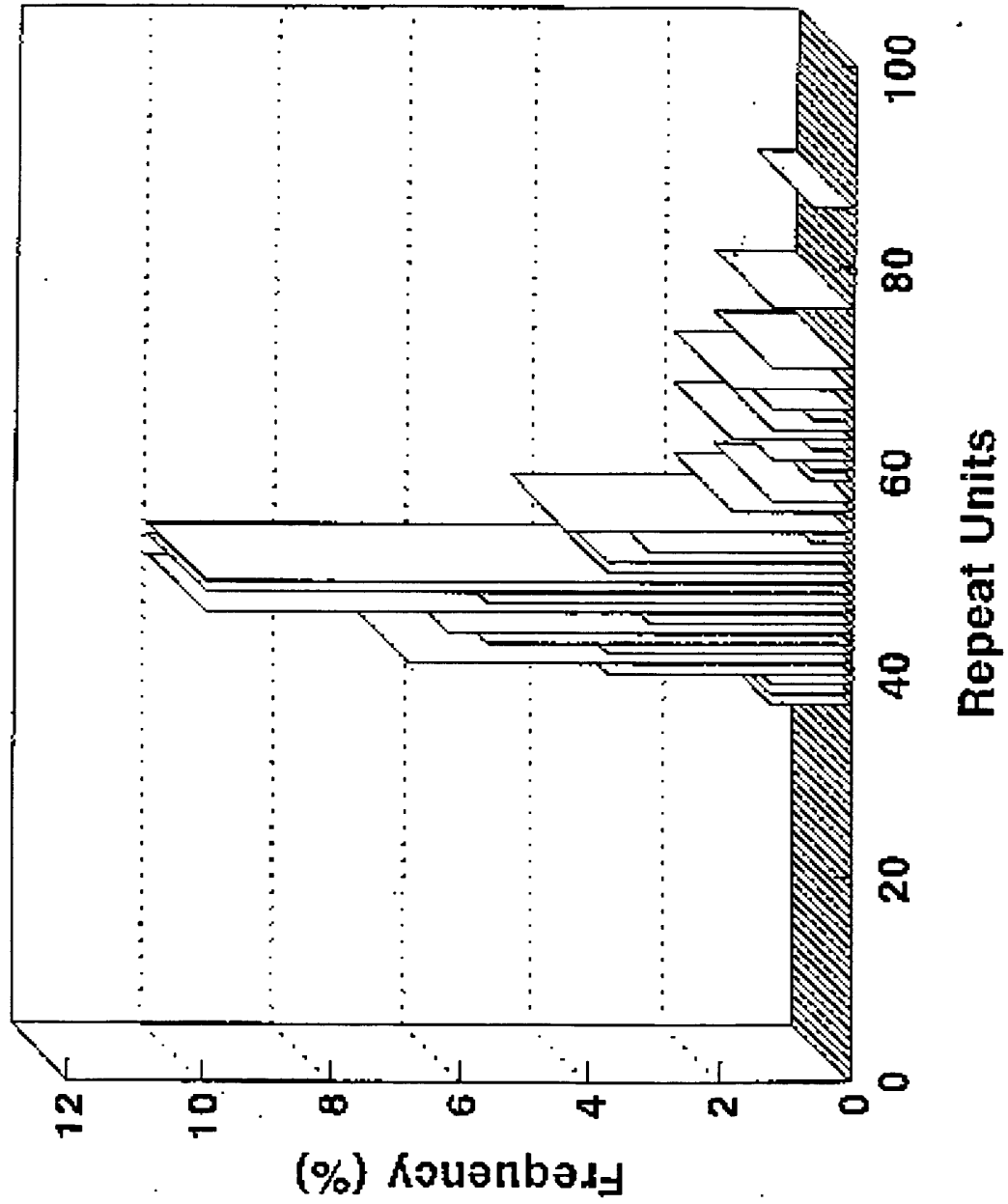


FIGURE 13 (PANEL A)

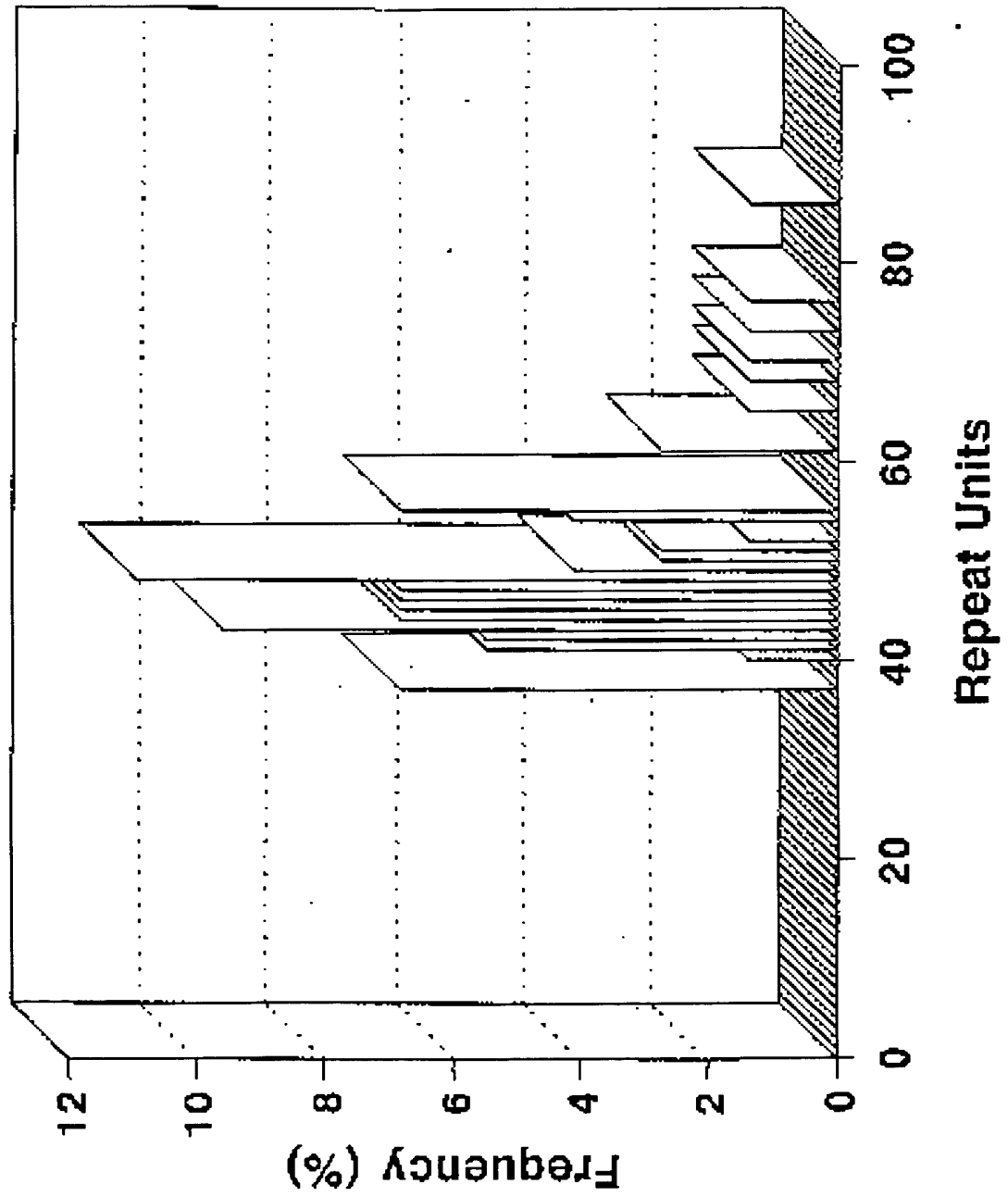


FIGURE 13 cont. (PANEL B)

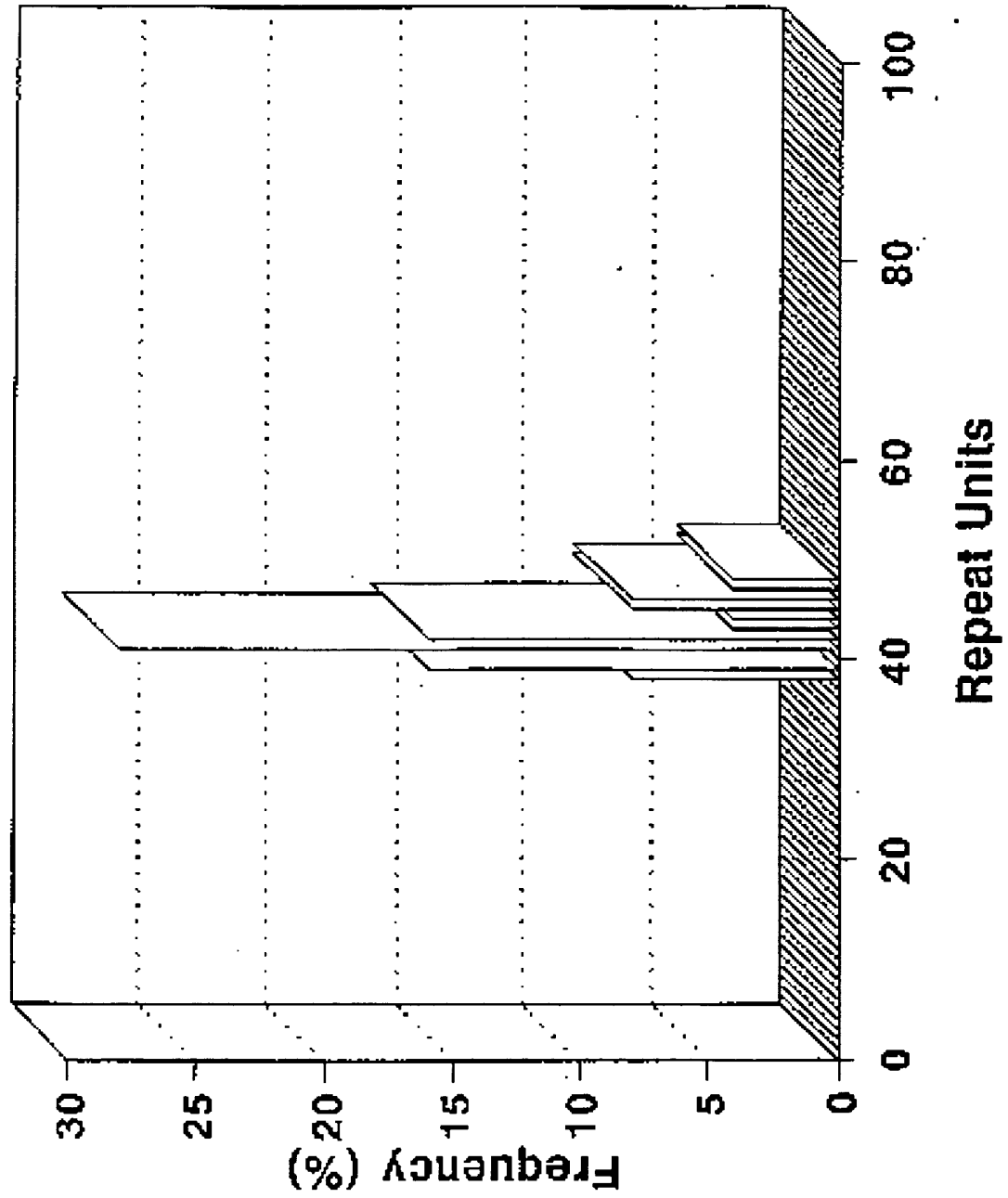


FIGURE 13 cont. (PANEL C)

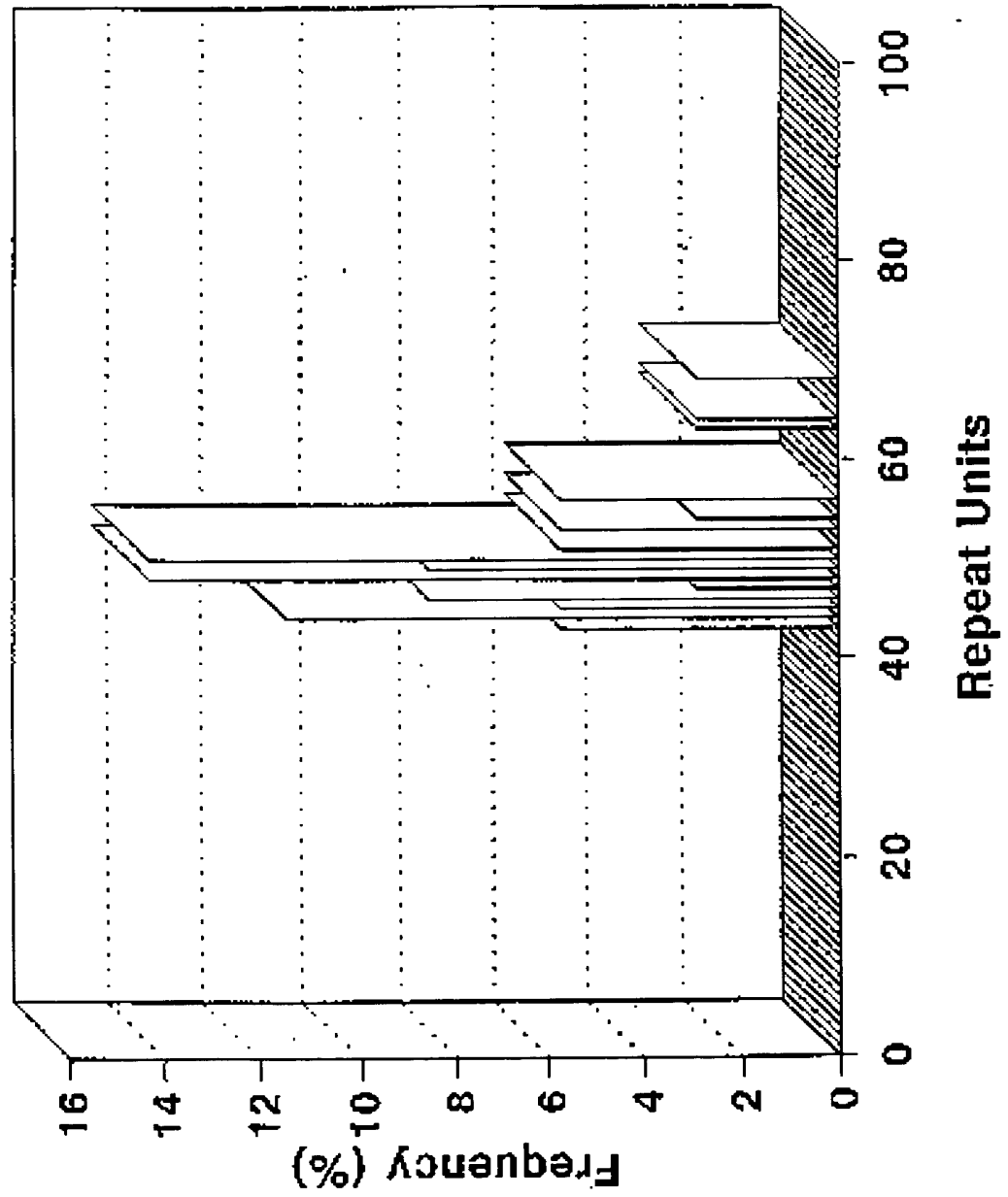
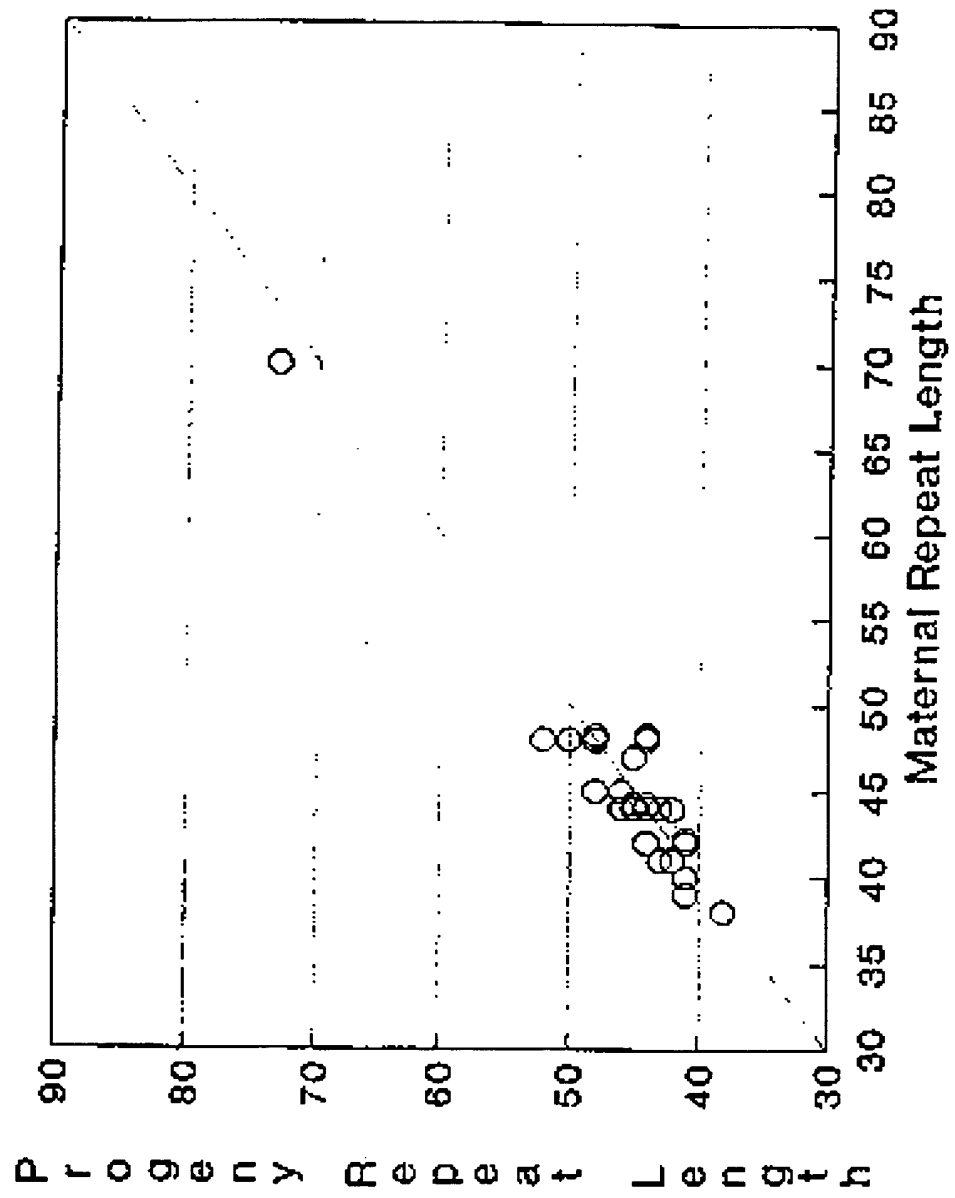


FIGURE 14 (PANEL A)



Scatter plot showing Progeny Repeat Length (Y-axis) versus Paternal Repeat Length (X-axis). The data points are represented by open circles. The X-axis ranges from 30 to 90, and the Y-axis ranges from 30 to 90. A dashed diagonal line represents the identity line (y=x). The data points generally fall below the identity line, indicating that the progeny repeat length is typically shorter than the paternal repeat length.

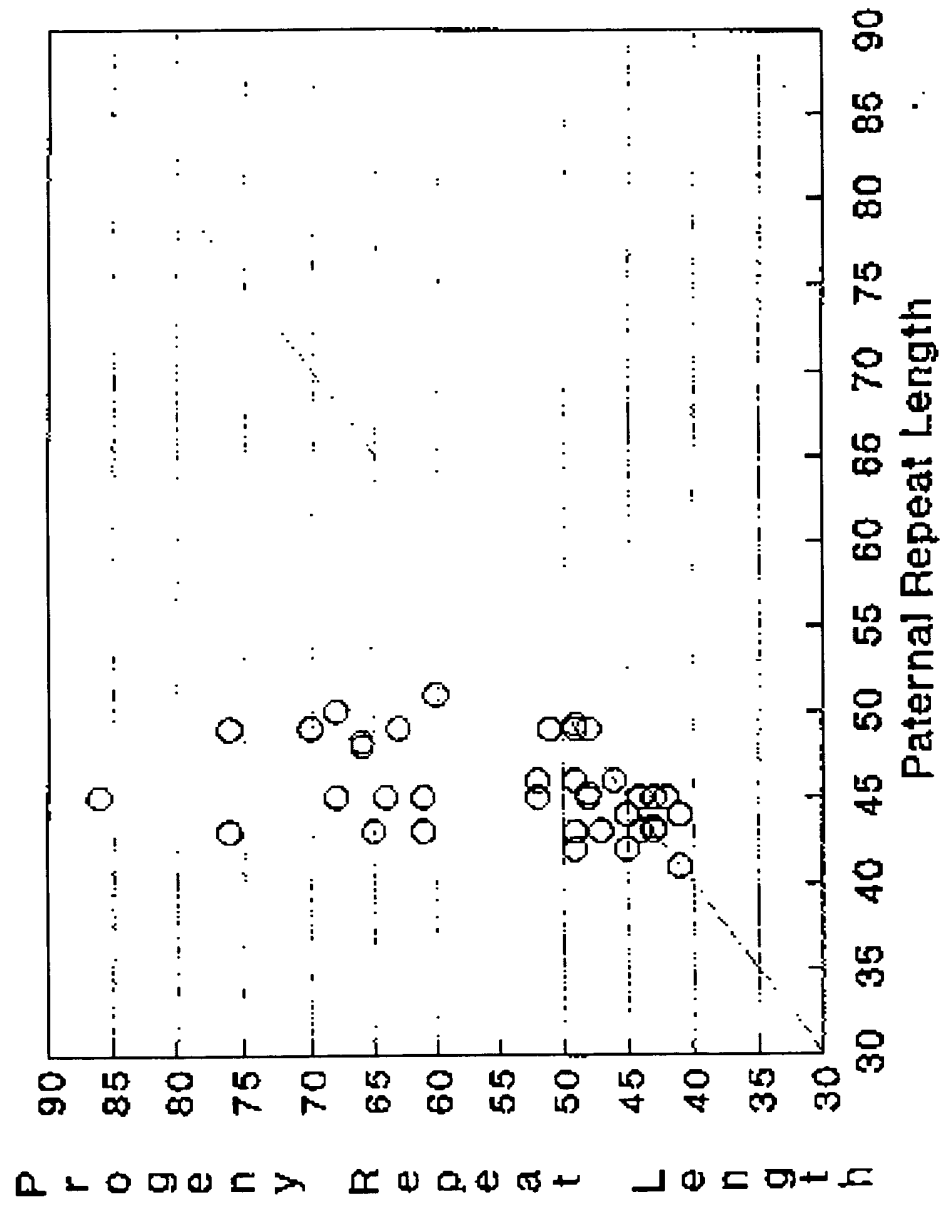


FIGURE 15

1	2	3	4	5
$\overline{\text{S L}}$	$\overline{\text{S L}}$	$\overline{\text{S L}}$	$\overline{\text{S L}}$	$\overline{\text{S L}}$

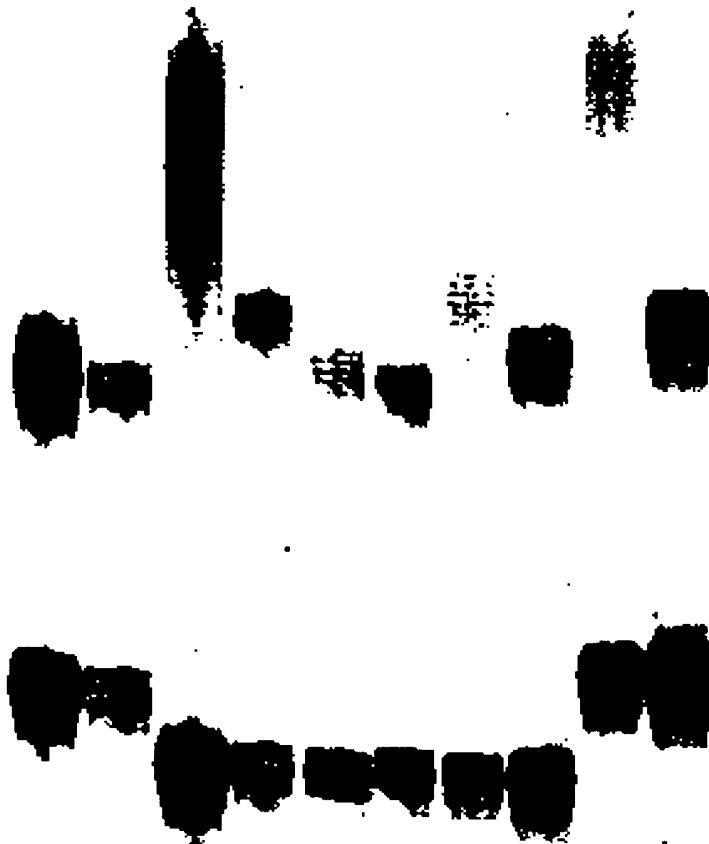
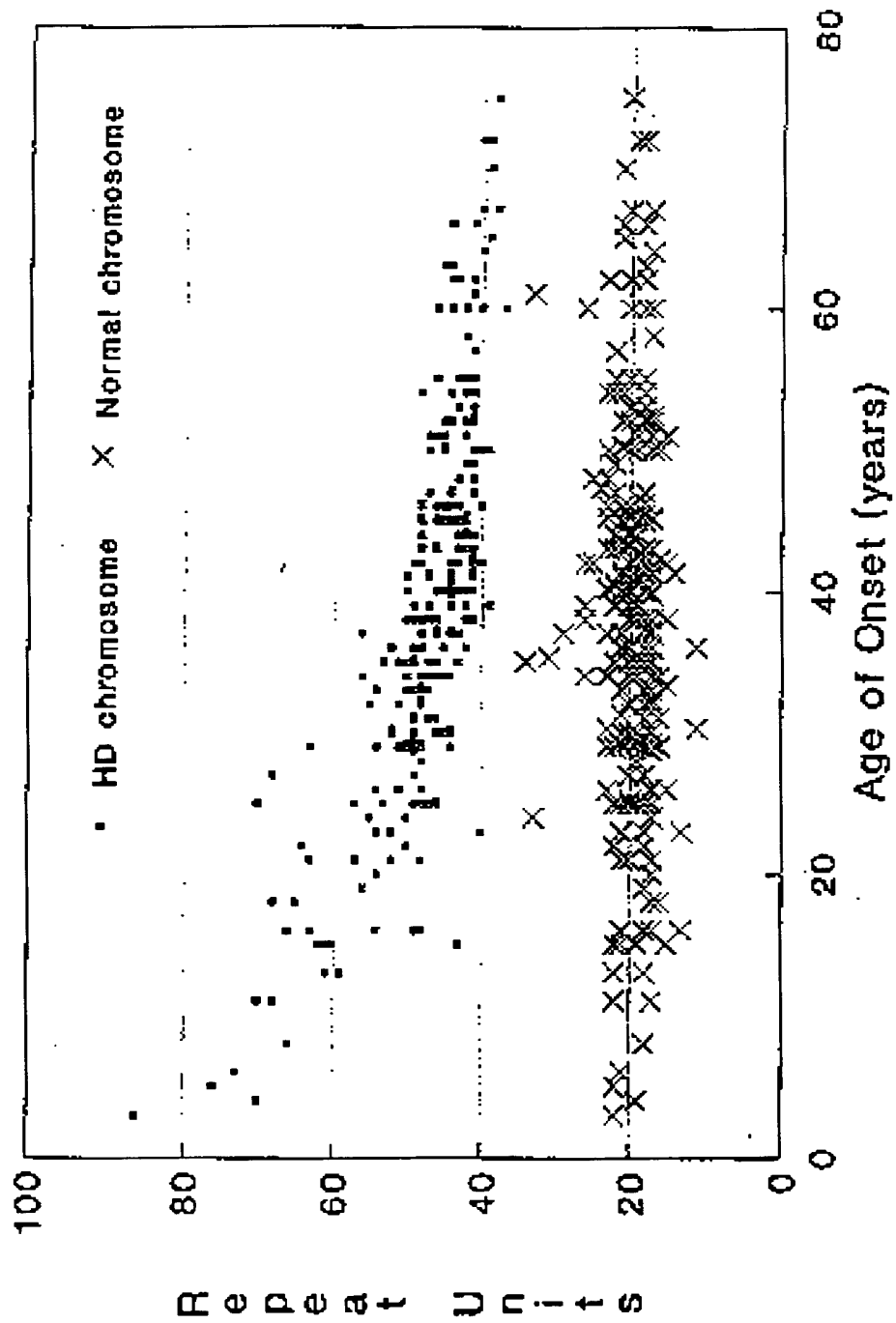
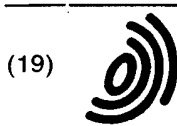


FIGURE 16





Europäisches Patentamt
European Patent Office
Office européen des brevets



(11) EP 0614 977 A3

(12) EUROPEAN PATENT APPLICATION

(88) Date of publication A3:
28.02.1996 Bulletin 1996/09

(43) Date of publication A2:
14.09.1994 Bulletin 1994/37

(21) Application number: 94301587.5

(22) Date of filing: 07.03.1994

(84) Designated Contracting States:
AT BE CH DE DK ES FR GB GR IE IT LI LU MC
NL PT SE

(30) Priority: 05.03.1993 US 27498
01.07.1993 US 85000

(71) Applicant:
THE GENERAL HOSPITAL CORPORATION
Boston, MA 02114 (US)

(51) Int Cl.⁶: C12N 15/12, C07K 13/00,
C12N 1/21, C12N 5/10,
C07K 15/28, C12N 5/16,
C12Q 1/68, A61K 37/02,
A61K 48/00, C12P 21/08

(72) Inventors:
• MacDonald, Marcy E.
Lexington, Massachusetts 02173 (US)
• Ambrose, Christine M.
Massachusetts 02129 (US)
• Duyao, Mabel P.
Cambridge, Massachusetts 02138 (US)
• Gusella, James F.
Framingham, Massachusetts 01701 (US)

(74) Representative: Wright, Simon Mark et al
London WC1N 2DD (GB)

(54) Huntingtin DNA, protein and uses thereof

(57) A novel gene, *huntingtin*, is described, encoding huntingtin protein, recombinant vectors and hosts capable of expressing huntingtin. Methods for the diagnosis and treatment of Huntington's disease are also provided.

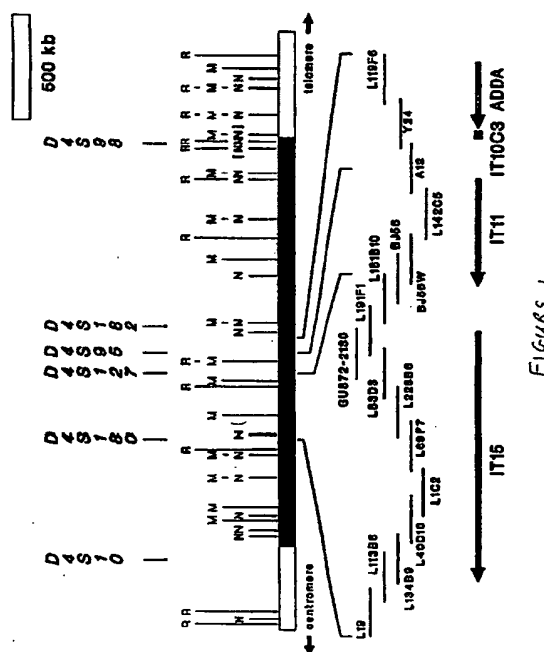


Figure 1



European Patent
Office

EUROPEAN SEARCH REPORT

Application Number
EP 94 30 1587

DOCUMENTS CONSIDERED TO BE RELEVANT			
Category	Citation of document with indication, where appropriate, of relevant passages	Relevant to claim	CLASSIFICATION OF THE APPLICATION (Int.Cl.5)
X,D	SOMAT. CELL MOL. GENET., vol. 17, no. 5, 1991 pages 481-488, LIN ET AL. 'New DNA markers in the Huntington's disease gene candidate region' * the whole document * ---	4,6	C12N15/12 C07K13/00 C12N1/21 C12N5/10 C07K15/28 C12N5/16 C12Q1/68 A61K37/02 A61K48/00 C12P21/08
X,D	NATURE GENET., vol. 1, May 1992 pages 99-103, MAC DONALD ET AL. 'The Huntington's disease candidate region exhibits many different haplotypes' * the whole document * ---	4,6	
P,X	CELL, vol. 72, 26 March 1993 pages 971-983, THE HUNTINGTON'S DISEASE COLLABORATIVE RESEARCH GROUP 'A novel gene containing a trinucleotide repeat that is expanded and unstable on Huntington's disease chromosomes' * the whole document * ---	1-6,8, 13-15	TECHNICAL FIELDS SEARCHED (Int.Cl.5) C07K C12N
P,X	CR ACAD. SCI. III, vol. 316, no. 11, November 1993 pages 1374-1380, DODÉ ET AL. 'Huntington's disease in French families : CAG repeat expansion and linkage disequilibrium analysis' * the whole document * ---	4,6, 13-15	
		-/--	
The present search report has been drawn up for all claims			
Place of search THE HAGUE		Date of completion of the search 15 December 1995	Examiner Gac, G
<p>CATEGORY OF CITED DOCUMENTS</p> <p>X : particularly relevant if taken alone Y : particularly relevant if combined with another document of the same category A : technological background O : non-written disclosure P : intermediate document</p> <p>T : theory or principle underlying the invention E : earlier patent document, but published on, or after the filing date D : document cited in the application L : document cited for other reasons * : member of the same patent family, corresponding document</p>			

EP 0 614 977 A3



European Patent
Office

EUROPEAN SEARCH REPORT

Application Number
EP 94 30 1587

DOCUMENTS CONSIDERED TO BE RELEVANT			
Category	Citation of document with indication, where appropriate, of relevant passages	Relevant to claim	CLASSIFICATION OF THE APPLICATION (Int.Cl.5)
P, X	MOL. CELL PROBES, vol. 7, no. 3, June 1993 pages 235-239, WARNER ET AL. 'A new polymerase chain reaction (PCR) assay for the trinucleotide repeat that is unstable and expanded on Huntington's disease chromosomes' * the whole document *	4, 6, 13-15	
A	US-A-4 666 828 (GUSELLA) 19 May 1987 * the whole document *	1-23	
A	MOL. CELL. BIOL., vol. 10, no. 11, November 1990 pages 5616-5625, LAURENT ET AL. 'The SNF5 protein of Saccharomyces cerevisiae is a glutamine- and proline-rich transcriptional activator that affects expression of a broad spectrum of genes' * page 5618 - page 5619 *	19	
			TECHNICAL FIELDS SEARCHED (Int.Cl.5)
The present search report has been drawn up for all claims			
Place of search THE HAGUE		Date of completion of the search 15 December 1995	Examiner Gac, G
<p>CATEGORY OF CITED DOCUMENTS</p> <p>X : particularly relevant if taken alone Y : particularly relevant if combined with another document of the same category A : technological background O : non-written disclosure P : intermediate document</p> <p>T : theory or principle underlying the invention E : earlier patent document, but published on, or after the filing date D : document cited in the application L : document cited for other reasons</p> <p>& : member of the same patent family, corresponding document</p>			

EPO FORM 1503 01.82 (P04C01)

